UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS

DE TELECOMUNICACIÓN

TESIS DOCTORAL

Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano

JUAN MANUEL MONTERO MARTÍNEZ

Ingeniero de Telecomunicación

Madrid, 2003

UNIVERSIDAD POLITÉCNICA DE MADRID

DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS

DE TELECOMUNICACIÓN

TESIS DOCTORAL

Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano

JUAN MANUEL MONTERO MARTÍNEZ

Ingeniero de Telecomunicación

Director de la Tesis

JOSÉ MANUEL PARDO MUÑOZ

Doctor Ingeniero de Telecomunicación

2003

Tesis Doctoral: Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano

Autor: JUAN MANUEL MONTERO MARTÍNEZ

Director: Dr. INGENIERO JOSÉ MANUEL PARDO MUÑOZ

El tribunal nombrado para juzgar la Tesis Doctoral arriba citada, compuesto por los doctores:

PRESIDENTE: *Dr. Javier Ferreiros López*

VOCALES: *Dr. Eduardo Rodríguez Banga*

Dr. Emilia Victoria Enríquez Carrasco

Dr. David Escudero Mancebo

SECRETARIO: *Dr. Ricardo de Córdoba Herralde*

acuerda otorgarle la calificación de

Sobresaliente cum Laude

Madrid, 14 de Noviembre de 2003

El Secretario del Tribunal

Agradecimientos

Al director de este trabajo, José Manuel Pardo, por todo el apoyo y la confianza que ha depositado en mí durante estos años, así como por ofrecerme la oportunidad de incorporarme al mundo de la investigación y de la docencia.

A todas las personas que han sido o son miembros del Grupo de Tecnología del Habla al que pertenezco, por su gran calidad profesional y personal: Javier Macías, Javier Ferreiros, José Colás, José David Romeral, Silvia Muñoz..., y muy especialmente a Ascensión Gallardo, a Rubén San Segundo y a Juana Gutiérrez-Arriola, y a los que han formado parte del grupo de síntesis en el que encuadra esta Tesis: Ricardo de Córdoba, José Ángel Vallejo, Mª Ángeles Romero, Emilia Enríquez y Francisco Giménez de los Galanes.

También quiero agradecer su colaboración a todos los alumnos a los cuales he dirigido su proyecto fin de carrera desde Sira hasta Jesús, pero especialmente a Gerardo Martínez Salas, Azucena Jiménez, Daniel Polanco, Rogelio Vargas, Julio Sánchez y Carlos Martín.

Gracias a Eduardo Jover, nuestro "emocionado" locutor, y a Johan Bertenstam y Kjell Gustafsson de KTH, que me ayudaron a dar mis primeros pasos en el mundo de la voz con emociones, y a Francisco Martínez-Sánchez y al personal de la empresa Natural Vox con los que tanto he colaborado.

Gracias, también, por su ayuda y apoyo al resto de miembros del Departamento de Ingeniería Electrónica, en especial a Ignacio Izpura, Fernando González Sanz, Mariano González Bédmar, y a todos los que colaboran en el buen funcionamiento de los laboratorios docentes.

A Juan Ramón, Fernando, Juan Ignacio, Rafa, Maria José, Mari Cruz..., a los que últimamente apenas he podido ver.

A Jesús Gomeza, a Maxi y compañía, porque sólo estamos lejos en la distancia.

A mis padres y a mi hermana y a toda mi familia paterna y materna, a Asen y a su familia, que llevan años aguantándome a pesar de tantas cosas...

A todos aquellos de los que el tiempo y las circunstancias me han alejado, aunque no definitivamente.

Resumen

En esta Tesis se abordan tres subproblemas relacionados con la variedad y la naturalidad en la conversión texto habla en castellano: el procesado lingüístico orientado a prosodia, el modelado de la frecuencia fundamental en un dominio restringido y el análisis, modelado y conversión texto a voz con emociones. El capítulo del estado de la cuestión recoge con detalle los principales progresos en cada módulo de un conversor. El primer apartado destacable está dedicado al análisis gramatical y sintáctico, cubriendo las técnicas de normalización del texto, los corpora anotados, las bases de datos léxicas disponibles en castellano, las técnicas de desambiguación contextual y de análisis sintáctico y los sistemas disponibles en castellano. En cuanto al modelado prosódico, se tratan los modelos empleados tanto para la frecuencia fundamental como el ritmo, las duraciones y el pausado, las principales escuelas de análisis de la curva de frecuencia fundamental y las técnicas avanzadas de diseño de las bases de datos. En el apartado dedicado a la voz emotiva se describen y comentan los principales sistemas internacionales desarrollados y las bases de datos disponibles. Como en general la síntesis por formantes ha dominado este campo, se describe esta técnica, para finalizar con una revisión de las alternativas de evaluación empleadas en síntesis de voz con emociones.

En el capítulo dedicado a las investigaciones en procesado lingüístico del texto se comienza describiendo en detalle los corpora empleado en la experimentación, tanto en normalización como en etiquetado. La técnica desarrollada en normalización emplea reglas de experto, con muy buenos resultados tanto en precisión como en cobertura, destacando el empleo de reglas de silabicación para la detección precisa de palabras extranjeras. Al afrontar la desambiguación gramatical, se comparan tres técnicas: reglas de experto, aprendizaje automático de reglas y modelado estocástico, obteniéndose los mejores resultados con esta última técnica, debido a su capacidad de procesar más adecuadamente textos fuera del dominio de entrenamiento. Finalmente se aborda el análisis sintáctico por medio de gramática de contexto libre como un proceso en dos fases:, una primera sintagmática y una segunda relacional básica, a fin de maximizar la cobertura del análisis. Para la resolución de las ambigüedades que nos permiten alcanzar gran cobertura se adapta el principio de mínima longitud de descripción con notables resultados. Las gramáticas desarrolladas se encuentran comentadas y ejemplificadas en un apéndice.

Para el modelado de F0 en un dominio restringido se emplean perceptrones multicapa. En una primera etapa se describe y evalúa una nueva técnica de diseño de base de datos basada en un algoritmo voraz moderado mediante subobjetivos intermedios. La exhaustiva experimentación con los diversos parámetros de predicción, la configuración de la red y las subdivisiones de la base de datos ocupa la mayor parte del capítulo, destacando la aportación de un parámetro específico del dominio restringido (el número de la frase portadora del texto que sintetizar) junto a otros más clásicos (acentuación, tipo de grupo fónico y posición en el mismo).

El capítulo dedicado a la voz emotiva comienza detallando el proceso de creación de una nueva voz castellana masculina en síntesis por formantes con modelo mejorado de fuente (reglas y metodología), evaluando las posibilidades de personalización de voz que ofrece. Para trabajar con voz con emociones se diseña, graba y etiqueta una base de datos de voz en la que un actor simula tristeza, alegría, sorpresa, enfado y también una voz neutra. Por medio de técnicas paramétricas (modelo de picos y valles en tono, y multiplicativo en las duraciones) se analiza prosódicamente la base de datos y se establece una primera caracterización de la voz en las distintas emociones. Empleando como base la voz personalizable se desarrolla el sistema completo de conversión texto a voz con emociones y se evalúa, destacando la rápida adaptación de los usuarios en cuanto a la identificación de la emoción expresada. Finalmente se experimenta con síntesis por concatenación y síntesis por copia, llegando a las siguientes conclusiones: la voz sorprendida se identifica prosódicamente, las características segmentales son las que caracterizan al enfado en frío; y, finalmente, la tristeza y la alegría son de naturaleza mixta.

Abstract

This doctoral Thesis studies three approaches in order to improve naturalness in text-to-speech conversion:

- **Linguistic processing**:
  - **Preprocessing**: for the normalization of the input text, I have developed and evaluated the use of a set of dictionaries and expert rules, getting a 85% precision on an evaluation corpus from a newspaper domain.

- **Lexical information**: using general dictionaries (not adapted to the evaluation domain), we got a 99,87% recall, comparable to the best systems for Spanish, beating the results of a stochastic system (with greater precision).

- **Automatic POS tagging**: I have tried 2 approaches for contextual disambiguation: automatic rule learning and stochastic modeling; the second approach achieves higher recall rates, specially for out-of-domain tests.

- **Shallow parsing**: adapting a context-free grammar system, I have developed and evaluated a new robust general-purpose grammar for shallow parsing (97%), using cut-rules for reducing the number of possible analyses, applying concordance rules as a filter and using the minimum number of simple segments in each analysis in order to get the best one. Using other context-free rules, we modeled the relations between several simple segments in order to build more complex ones.

§ **Restricted-domain F0 modeling**:

- **New greedy algorithm for database design**, capable of summing up a big database with a precision higher than 95%, taking into account several prosodic and segmental feature vectors.

- **F0 modeling on a restricted domain**, using a multilayer perceptron, with new parameters such as the number of the carrier sentence, and the analysis on how to group the recordings in order to get the best possible modeling.

§ **Analysis, modeling and emotional TTS conversion**:

- **Development of a new configurable formant-based voice in Spanish**, including the evaluation of the adaptation process.

- **Design and recording of the first emotional speech database in Spanish**, design for its use in prosody synthesis; analysis of the prosody using parametric techniques and its evaluation in copy-synthesis experiments.

- **Development of the first emotional formant-based Spanish synthesizer**, and its evaluation.

- I have analysed **whether the segmental properties or the prosodic properties make identifiable the simulated emotion**: cold anger is detectable through its segmental characteristics; surprise is detected through pitch and tempo; for joy and sadness both segments and prosody are necessary.

Índice

Capítulo 1 **Introducción**

1.1 Objetivos de la Tesis

1.1.1 *Procesado lingüístico automático*

1.1.2 *Modelado de la F0 para síntesis en dominio restringido*

1.1.3 *Análisis y síntesis de habla con emociones*

1.2 Contenido de la Tesis

Capítulo 2 **Estado de la cuestión**

2.1 Introducción

2.1.1 *Sistemas comerciales de conversión texto a voz*
2.2 Procesado lingüístico
2.2.1 *Etiquetado morfosintáctico automático*
2.2.1.1 Teoría lingüística generativa
2.2.1.2 Preprocesamiento
2.2.1.3 Diccionarios y plataformas léxicas en castellano
2.2.1.4 Técnicas de desambiguación en el etiquetado morfosintáctico
2.2.1.5 Sistemas combinados o integrados
2.2.1.6 Medidas de evaluación y comparación entre sistemas
2.2.1.7 Etiquetado manual
2.2.1.8 Corpora en castellano
2.2.1.9 Sistemas de desambiguación en castellano
2.2.2 *Sintaxis y análisis sintagmático*
2.2.2.1 Características de los segmentos o sintagmas simples
2.2.2.2 Sistemas automáticos de segmentación en sintagmas simples
2.2.2.3 Corpus y bases de datos sintácticos en castellano
2.2.2.4 Sistemas de análisis sintáctico en castellano
2.3 Análisis y modelado prosódico
2.3.1 *Entonación y F0*
2.3.1.1 Escuelas de análisis de contornos de F0
2.3.1.2 Acentuación y desacentuación léxica. Foco
2.3.1.3 Relaciones entre F0, intensidad y duración
2.3.1.4 Micro-prosodia o micro-melodía
2.3.1.5 Relaciones entre entonación y sintaxis
2.3.1.6 Patrones entonativos en castellano
2.3.1.7 Definición y diseño de una base de datos prosódica
2.3.1.8 Métodos para la generación de curvas de F0
2.3.1.9 Percepción de la frecuencia fundamental
2.3.1.10 Normalización de valores de F0
2.3.1.11 Evaluación del modelado de F0
2.3.2 *Duración y ritmo*
2.3.2.1 Normalización de la duración
2.3.2.2 Modelos de duraciones
2.3.3 *Pausado*
2.4 Personalización de voz y habla con emociones
2.4.1 *Síntesis por formantes*
2.4.2 *Sistemas de síntesis de voz con emociones*
2.4.2.1 El sistema Affect Editor
2.4.2.2 El sintetizador Hamlet
2.4.3 *Prótesis vocales*
2.4.4 *Bases de datos de voz con emociones*
2.4.4.1 Bases de datos en castellano
2.4.5 *Evaluación de sistemas de voz con emociones*

Capítulo 3 Procesado lingüístico automático

3.1 Introducción
3.2 Etiquetado morfosintáctico automático
3.2.1 *Corpora empleados*
3.2.1.1 El corpus de El Mundo
3.2.1.2 El corpus 860
3.2.2 *Modelado léxico*
3.2.2.1 Normalizador
3.2.2.2 Diccionarios
3.2.2.3 Conjugador verbal
3.2.2.4 Reglas léxicas externas o de terminaciones
3.2.2.5 Cobertura léxica
3.2.3 *Desambiguación contextual*
3.2.3.1 Creación de reglas manuales contextuales
3.2.3.2 Aprendizaje automático de reglas
3.2.3.3 Desambiguación contextual estocástica
3.2.4 *Conclusiones sobre etiquetado automático*
3.3 Análisis sintáctico automático y robusto
3.3.1 *Análisis sintáctico*
3.3.2 *El algoritmo CYK*
3.3.2.1 Recuperación de todos los análisis correctos
3.3.3 *Texto categorizado y reglas léxicas*
3.3.4 *Análisis sintagmático y reglas de corte*
3.3.4.1 Resultados
3.3.5 *Reglas gramaticales sintagmáticas*
3.3.5.1 Principales segmentos (sintagmas simples)
3.3.5.2 Secuencia de segmentos (sintagmas simples)
3.3.5.3 Filtros de concordancia
3.3.5.4 Principio de mínima longitud de la descripción
3.3.5.5 Evaluación
3.3.5.6 Recategorización
3.3.6 *Reglas gramaticales de segundo nivel (sintácticas)*
3.3.6.1 Evaluación
3.3.7 *Conclusiones sobre análisis sintáctico*

Capítulo 4 Modelado de la F0 para síntesis en dominio restringido

4.1 Diseño de la base de datos de dominio restringido
4.1.1 *Criterios de selección del contenido de los campos variables*
4.1.2 *Simplificación de los criterios*
4.1.2.1 Simplificación para la base de datos de nombres propios
4.1.2.2 Simplificación para la base de datos con sintagmas nominales en oraciones enunciativas
4.1.2.3 Simplificación para la base de datos con sintagmas nominales en oraciones interrogativas
4.1.3 *Algoritmo de selección*
4.1.4 *Resultados*
4.1.4.1 Ejemplo de selección de 100 pueblos

4.1.4.2 Ejemplo de selección de 150 pueblos
4.1.4.3 Ejemplo de selección de 250 pueblos
4.1.4.4 Ejemplo de selección de 150 apellidos
4.1.4.5 Ejemplo de selección de 60 apellidos
4.1.4.6 Ejemplos de selección con baja ratio de ejemplos disponibles
4.1.4.7 Errores graves de selección
4.1.4.8 Algoritmo con subobjetivos intermedios
4.2 Grabación y etiquetado de la base de datos
4.3 Análisis y parametrización
4.4 Condiciones generales de experimentación para el modelado de F0 mediante redes neuronales artificiales
4.4.1 Consideraciones generales
4.4.2 Parámetros que se ensayarán
4.4.2.1 Nuevas codificaciones de Inicial, Acentuada y Final
4.4.2.2 Nuevos parámetros o elementos de parametrización
4.4.3 Elementos relacionados con la propia red
4.4.4 Organización de la experimentación
4.4.5 Estrategia de experimentación
4.5 Experimentos sobre nombres propios en enunciativas
4.5.1 Experimento de base con nombres propios
4.5.2 Experimentos sobre la influencia de la eliminación del zscore en el experimento de base de nombres propios
4.5.3 Experimentos sobre la influencia de no codificar la información sobre sílabas iniciales, finales o acentuadas en el experimento de base de nombres propios
4.5.3.1 Omisión del elemento 'sílaba inicial'
4.5.3.2 Omisión del elemento 'sílaba acentuada'
4.5.3.3 Omisión del elemento 'sílaba final'
4.5.3.4 Experimento de base de nombres propios omitiendo varios elementos (sílaba inicial, sílaba acentuada o sílaba final)
4.5.4 Experimentos sobre la influencia de eliminar el elemento 'signo de puntuación final' en el experimento de base de nombres propios
4.5.5 Experimentos sobre la influencia de codificar el número de sílabas en el experimento de base de nombres propios
4.5.6 Segundo experimento de base de nombres propios: influencia de codificar el número de frase portadora
4.5.7 Experimentos de nombres propios sobre otros parámetros
4.5.8 Conclusiones sobre el modelado de nombres propios en enunciativas
4.6 Experimentos sobre frases interrogativas
4.6.1 Experimentos de base de interrogativas
4.6.2 Experimentos sobre la influencia de la no codificación del número de la frase portadora en el experimento de base de interrogativas
4.6.3 Experimentos sobre la influencia de otros parámetros en el experimento de base de interrogativas
4.6.4 Conclusiones sobre el modelado de interrogativas
4.7 Experimentos sobre frases enunciativas con sintagmas nominales largos
4.7.1 Experimentos de base de sintagmas nominales
4.7.2 Experimentos sobre la influencia de la no inclusión del elemento 'signo de puntuación final'
4.7.3 Experimentos sobre la no codificación del número de la frase portadora
4.7.4 Experimentos sobre la no codificación del número de sílabas
4.7.5 Experimentos sobre otros parámetros
4.7.6 Conclusiones sobre enunciativas

4.8 Experimentos con las frases especiales
4.8.1 Condiciones de experimentación
4.8.2 Experimentos con las frases especiales 6 y 7
4.8.2.1 Experimentos conjuntos con las frases 6 y 7
4.8.2.2 Experimentos con la frase especial 6
4.8.2.3 Experimentos con la frase especial 7
4.8.2.4 Experimentos con las frases especiales 6 y 7 agrupadas con los demás nombres propios
4.8.2.5 Conclusiones sobre las frases 6 y 7
4.8.3 Experimentos con la frase especial 8
4.8.4 Experimentos con la frase especial 15
4.8.4.1 Experimentos con la frase especial 15 considerada como interrogativa
4.8.4.2 Experimentos con la frase especial 15 considerada como enunciativa
4.9 Experimento global conjunto con todas las frases
4.10 Conclusiones sobre el modelado de F0 en dominio restringido

Capítulo 5 Análisis y síntesis de habla con emociones

5.1 Desarrollo de una nueva voz personalizable mediante síntesis por formantes
5.2 Evaluación de la voz personalizada y del proceso de personalización
5.2.1 Descripción de las sesiones de trabajo para la evaluación del proceso de personalización
5.2.2 Resultados
5.2.2.1 Valores personalizados de los parámetros para cada usuario
5.2.3 Evaluación de la calidad global de la voz sintética
5.2.3.1 ¿Cómo de natural suena la voz?
5.2.3.2 ¿Cómo es de inteligible el habla?
5.2.3.3 ¿Cómo calificaría la calidad de la voz?
5.3 La base de datos SES: *Spanish Emotional Speech*
5.3.1 Frases cortas
5.3.2 Palabras aisladas
5.3.3 Párrafos de corta longitud
5.3.4 Grabación
5.3.5 Etiquetado y marcado de SES
5.3.6 Análisis de SES
5.3.6.1 Análisis cualitativo
5.3.6.2 Análisis cuantitativo de las duraciones y el ritmo
5.3.6.3 Análisis cuantitativo de la entonación
5.3.6.4 Síntesis por formantes de voz con emociones
5.4 Evaluación del habla con emociones empleando síntesis por formantes
5.4.1 Parámetros generales de la evaluación
5.4.1.1 Estímulos
5.4.2 Sesiones de trabajo con los oyentes
5.4.3 Resultados
5.4.3.1 Identificación de la emoción transmitida por la voz sintética
5.4.3.2 Matrices de confusión para voz sintética
5.4.3.3 Resultados totales de reconocimiento de la emoción simulada

5.4.3.4 Resultados para las 10 primeras grabaciones
5.4.3.5 Resultados para las 10 últimas grabaciones
5.4.3.6 Resultados para voz natural
5.4.3.7 Matrices de confusión para voz natural
5.4.3.8 Identificación de la emoción simulada en función del número de frase
5.5 Conclusiones sobre síntesis de voz con emociones mediante síntesis por formantes
5.6 Experimentos de síntesis-por-copia y voz con emociones
5.6.1 Conclusiones sobre síntesis de voz con emociones mediante síntesis por copia

Capítulo 6 Conclusiones y líneas futuras

6.1 Conclusiones
6.1.1 Procesado lingüístico automático
6.1.2 Modelado de F0 en dominio restringido
6.1.3 Análisis y síntesis de voz con emociones
6.2 Líneas futuras
6.2.1 Procesado lingüístico automático
6.2.1.1 Categorización automática
6.2.1.2 Análisis sintáctico
6.2.1.3 Análisis semántico
6.2.2 Modelado de F0 en dominio restringido
6.2.2.1 Nuevo método voraz para diseño de bases de datos
6.2.2.2 Modelado de F0
6.2.3 Análisis y síntesis de voz con emociones
6.2.3.1 Síntesis de voz configurable y con emociones
6.2.3.2 Base de datos de habla emotiva en castellano

Referencias

Apéndices

A.1 Procesado lingüístico automático
A.1.1 Etiquetado del 860
A.1.1.1 Nuevo etiquetado del corpus 860
A.1.1.2 Formato de las etiquetas del 860
A.1.1.3 Categorías primarias y secundarias
A.1.2 Lista de paradigmas irregulares empleados
A.1.3 Patrones del experimento de aprendizaje de reglas de categorización
A.1.4 Conjuntos de etiquetas del experimento de aprendizaje de reglas de categorización
A.1.5 Tablas de resultados de los experimentos sobre etiquetado estocástico
A.1.6 Reglas léxicas de preprocesamiento para el análisis sintáctico
A.1.7 Gramáticas de contexto libre empleadas
A.1.8 Gramática de primer nivel
A.1.8.1 Secuencia de segmentos o sintagmas simples
A.1.8.2 Nexos
A.1.8.3 Formas verbales
A.1.8.4 Nombres propios
A.1.8.5 Sintagma nominal

A.1.8.6 Sintagma adverbial
A.1.8.7 Sintagma adjetival
A.1.8.8 Estructuras con determinante
A.1.8.9 Sintagmas preposicionales
A.1.8.10 Locuciones
A.1.9 Gramática de segundo nivel
A.1.9.1 Cuantificación
A.1.9.2 Fechas
A.1.9.3 Comparaciones
A.1.9.4 Coordinación
A.1.9.5 Comillas
A.2 Modelado de F0 en dominio restringido
A.2.1 Frases patrón iniciales de la base de datos de dominio restringido
A.2.2 Frases patrón definitivas de la base de datos de dominio restringido
A.2.3 Análisis estadístico del modelado de F0 parámetro a parámetro
A.2.4 Análisis de F0 con un modelo paramétrico en dominio restringido
A.2.4.1 Nombres propios en enunciativas
A.2.4.2 Sintagmas nominales en enunciativas
A.3 Análisis y síntesis de habla con emociones
A.3.1 Personalización de voz
A.3.1.1 Evaluación inicial del sintetizador (previo a la personalización)
A.3.1.2 Bases de datos para una voz neutra
A.3.1.3 Herramientas semiautomáticas
A.3.1.4 Diseño e implementación de una nueva voz
A.3.1.5 Reglas Prosódicas
A.3.1.6 Nuevas reglas Segmentales
A.3.1.7 Integración y pruebas
A.3.2 Ejemplo de cuestionario para la evaluación de síntesis de voz con emociones
A.3.3 Textos de la base de datos SES
A.3.3.1 Párrafos
A.3.3.2 Frases
A.3.3.3 Palabras
6.2.3.3 Relación entre las frases y las palabras de la base de datos
A.3.4 Cuestionario de evaluación de voz emotiva en el proyecto VAESS
A.3.5 Cuestionario sobre la personalización de voz
A.3.6 Definición de rasgos simples y complejos para la voz personalizada o con emociones
A.3.7 Reglas segmentales y de entonación para el castellano (para personalización y para emociones)

Índice de tablas, cuadros e ilustraciones

Tabla 1 Principales parámetros del corpus 860

Tabla 2 Parámetros secundarios del corpus 860

Tabla 3 Comparación entre la distribución de los distintos símbolos (en tanto por uno) en los corpora de entrenamiento, de evaluación y el completo: **V**erbo, **N**ombre sustantivo, **A**djetivo, adver**B**io, p**R**onombre, **P**reposición, **D**eterminante, **C**onjunción, **I**nterfección, **M**iscelanea y otros (**L**)

Tabla 4 Resultados de imprecisión para varios tipos de palabra no normalizadas, según la información léxica empleada en su detección

Tabla 5 Resumen de los diccionarios que se emplean en el modelado léxico

Tabla 6 Resultados de etiquetado automático sin desambiguación contextual

Tabla 7 Resultados de cobertura léxica del etiquetador automático TnT

Tabla 8 Resultados de la evaluación con aprendizaje de reglas

Tabla 9 Gráfica de cobertura con el conjunto de etiquetas completo, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando unigramas

Tabla 10 Gráfica de cobertura con el conjunto de etiquetas completo, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando unigramas

Tabla 11 Gráfica de cobertura con el conjunto de etiquetas simplificadas, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando unigramas

Tabla 12 Gráfica de cobertura con el conjunto de etiquetas simplificadas, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando unigramas

Tabla 13 Gráfica de cobertura con el conjunto de etiquetas simplificadas, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando bigramas

Tabla 14 Gráfica de cobertura con el conjunto de etiquetas simplificadas, con procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando bigramas

Tabla 15 Gráfica de cobertura con el conjunto de etiquetas completo, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.130 palabras, empleando bigramas

Tabla 16 Gráfica de cobertura con el conjunto de etiquetas completo, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando bigramas

Tabla 17 Gráfica de cobertura con el conjunto de etiquetas simplificado, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando trigramas

Tabla 18 Gráfica de cobertura con el conjunto de etiquetas simplificado, con procesado especial de locuciones, sobre un conjunto de evaluación del dominio de discapacidad de 22.518 palabras, empleando bigramas

Tabla 19 Gráfica de cobertura con el conjunto de etiquetas simplificado, con procesado especial de locuciones, sobre un conjunto de evaluación del dominio de discapacidad de 22.518 palabras, empleando trigramas

Cuadro 1 Algoritmo CYK

Cuadro 2 Algoritmo de reconstrucción CYK

Ilustración 1: Esquema de los principales sintagmas analizados y sus relaciones

Cuadro 3 Algoritmo voraz de selección de ejemplos para una base de datos de voz

Tabla 20 Resultados de selección de 100 pueblos

Tabla 21 Errores de selección de 100 pueblos

Tabla 22 Resultados de selección de 150 pueblos

Tabla 23 Errores de selección de 150 pueblos

Tabla 24 Resultados de selección de 250 pueblos

Tabla 25 Resultados de selección de 150 apellidos

Tabla 26 Resultados de selección de 60 apellidos

Tabla 27 Resultados de selección de 60 bancos

Tabla 28 Resultados de selección de 150 puertos

Tabla 29 Resultados de selección graves (entre 100 y 5000 ejemplos de pueblos y apellidos)

Tabla 30 Resultados de selección de pueblos en 1 paso y en 10 pasos (entre 100 y 250 pueblos)

Tabla 31 Experimento de base de nombres propios, empleando zscore, los elementos sílaba inicial, sílaba final y sílaba acentuada

con diversos tamaños de la ventana del contexto, con la codificación 1 del número de sílabas y empleando 4 bits para codificar el signo de puntuación final del grupo fónico.

Tabla 32 Experimento de base de nombres propios sin zscore.

Tabla 33 Experimento de base de nombres propios, sin emplear el elemento 'sílaba inicial' (pero sí los elementos 'sílaba final' o 'sílaba acentuada').

Tabla 34 Experimento de base de nombres propios sin emplear 'sílaba acentuada' (pero sí los elementos sílaba inicial y sílaba final).

Tabla 35 Experimento de base de nombres propios, sin emplear el elemento 'sílaba final' (pero sí los elementos sílaba inicial y sílaba acentuada).

Tabla 36 Experimento de base de nombres propios, empleando el elemento final (no el de inicial ni el de acentuada), con 20 neuronas ocultas.

Tabla 37 Experimento de base de nombres propios, empleando el elemento inicial (no el de final o el de acentuada) con 20 neuronas en la capa oculta.

Tabla 38 Experimento de base de nombres propios, empleando el elemento acentuada (no el de final o el de inicial) con 20 neuronas en la capa oculta.

Tabla 39 Experimento de base de nombres propios sin emplear bits para codificar el elemento 'signo de puntuación final del grupo fónico', con 20 neuronas en la capa oculta.

Tabla 40 Experimento de base de nombres propios sin emplear el elemento 'número de sílabas', con 20 neuronas en la capa oculta.

Tabla 41 Experimento de base de nombres propios empleando codificación 2 para el número de sílabas, con 20 neuronas en la capa oculta.

Tabla 42 Segundo experimento de base de nombres propios: incluye además la codificación del número de frase portadora y 20 neuronas en la capa oculta.

Tabla 43 Segundo experimento de base de nombres propios empleando el parámetro "es final de palabra" codificado sin ventana (valor 1) o con ventana +-1 (valor 3).

Tabla 44 Segundo experimento de base de nombres propios con el elemento "número de palabras" codificado (valor 1) o no (valor 0).

Tabla 45 Segundo experimento de base de nombres propios con el elemento "palabra en posición final" codificado (valor 1) o no (valor 0).

Tabla 46 Segundo experimento de base de nombres propios con el elemento "es palabra función" codificado (valor 1) o no (valor 0).

Tabla 47 Experimento de base de interrogativas con zscore, empleando los elementos acentuada, inicial y final con tamaño de ventana del contexto igual a 1, con codificación del 1 número de sílabas, empleando 4 bits para codificar el signo de puntuación final del grupo fónico y con codificación del número de frase portadora.

Tabla 48 Experimento de base de interrogativas sin codificación del número de frase portadora y con 20 neuronas en la capa oculta.

Tabla 49 Experimento de base de interrogativas codificando si es final de palabra sin contexto (1) o con contexto +-1 (3) y con 20 neuronas en la capa oculta.

Tabla 50 Experimento de base de interrogativas codificando si es final de palabra sin contexto (1) o con contexto +-1 (3) y con 20 neuronas en la capa oculta.

Tabla 51 Experimento de base de interrogativas, codificando el número de palabras y con 20 neuronas en la capa oculta.

Tabla 52 Experimento de base de interrogativas, codificando o no la pertenencia a palabras función y con 20 neuronas en la capa oculta.

Tabla 53 Experimento de base de interrogativas, empleando o no 5 bits para codificar el signo de puntuación inicial del grupo fónico y con 20 neuronas en la capa oculta..

Tabla 54 Experimentos de base de sintagmas nominales, con zscore, empleando los elementos acentuada, inicial y final con tamaño de ventana del contexto entre 1 y 5, sin codificación del número de sílabas, empleando 4 bits para codificar el signo de puntuación final del grupo fónico, codificación del número de frase portadora, con 20 neuronas en la capa oculta.

Tabla 55 Experimentos de base de sintagmas nominales sin codificar el signo de puntuación final del grupo fónico.

Tabla 56 Experimentos de base de sintagmas nominales con (19) y sin (0) codificación del número de frase portadora.

Tabla 57 Experimentos de base de sintagmas nominales sin codificar el signo de puntuación final del grupo fónico, con (19) o sin (0) codificación del número de frase portadora.

Tabla 58 Experimentos de base de sintagmas nominales, con (1 o 2) o sin (0) codificación del número de sílabas.

Tabla 59 Experimentos de base de sintagmas nominales con codificación 2 del número de sílabas y con (1 o 3) codificación de la pertenencia a una palabra función.

Tabla 60 Mejor resultado del experimento de base de sintagmas nominales con codificación 2 del número de sílabas y con codificación del número de palabras.

Tabla 61 Mejor resultado del experimento de base de sintagmas nominales, con codificación 2 del número de sílabas, y con (1) o sin (0) emplear codificación del número de palabras.

Tabla 62 Mejores resultados del experimento de base de sintagmas nominales, con codificación 2 del número de sílabas, , codificando si es final de palabra.

Tabla 63 Mejor resultado del experimento de base de sintagmas nominales con codificación 2 del número de sílabas y con codificación del signo inicial de puntuación.

Tabla 64 Mejor resultado del experimento de base de sintagmas nominales con codificación 2 del número de sílabas, con codificación del signo de puntuación anterior, con codificación de la posición de la sílaba en el final de una palabra y con codificación 3 de la pertenencia a una palabra función.

Tabla 65 Experimento de base de sintagmas nominales con codificación 2 del número de sílabas, con (5) o sin (0) codificación del signo de puntuación anterior, con (1) o sin (0) codificación de la posición de la sílaba en el final de una palabra, con (1) o sin (0) codificación de la posición de la palabra en la frase, con (1) o sin (0) codificación de la pertenencia a palabra función.

Tabla 66 Experimentos con las frases 6 y7.

Tabla 67 Experimentos con la frase 6.

Tabla 68 Experimentos con la frase 7.

Tabla 69 Experimento agrupando las frases 6 y 7 con los nombres propios.

Tabla 70 Experimento con la frase especial 15 considerada como interrogativa.

Tabla 71 Experimento con la frase especial 15 considerada como enunciativa

Tabla 72 Experimento general conjunto con todas las frases

Tabla 73 Valores de los parámetros para cada usuario.

Ilustración 2 Fragmento de voz neutra (parte superior) y su correspondiente de voz enfadada (parte inferior).

Tabla 74 Variación de diversos parámetros de duración entre las distintas emociones.

Tabla 75 Ratio entre el modelado de duración en las frases y en los párrafos para las distintas emociones.

Tabla 76 Duración media de las pausas por signos de puntuación en las distintas emociones.

Tabla 77 Resultados del análisis cuantitativo de la entonación de las frases para las diversas emociones.

Tabla 78 Resultados del análisis cuantitativo de la entonación de los párrafos para las diversas emociones.

Tabla 79 Resultados de la evaluación de identificación de la emoción transmitida por la voz sintética.

Tabla 80 Resultados totales de identificación de la emoción simulada.

Tabla 81 Resultados de identificación de la emoción simulada para las 10 primeras grabaciones.

Tabla 82 Resultados de identificación de la emoción simulada para las 10 últimas grabaciones.

Tabla 83 Matriz de confusión para la voz natural.

Tabla 84 Resultados de identificación para la voz sintética y para la voz natural.

Tabla 85 Resultados de identificación de emociones generadas mediante el método de síntesis por copia.

Tabla 86 Resultados de identificación de emociones generadas mediante el método de síntesis por copia.

Tabla 87 Resultados de identificación de emociones generadas mediante re-síntesis automática de prosodia.

Tabla 88 Experimentos con el conjunto de etiquetas completo, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando unigramas

Tabla 89 Experimentos con el conjunto de etiquetas completo, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando unigramas

Tabla 90 Experimentos con el conjunto de etiquetas simplificadas, sin procesado especial de locuciones, sobre un conjunto de

evaluación de 38.310 palabras, empleando unigramas

Tabla 91 Experimentos con el conjunto de etiquetas simplificadas, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando unigramas

Tabla 92 Experimentos con el conjunto de etiquetas simplificadas, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando bigramas

Tabla 93 Experimentos con el conjunto de etiquetas simplificadas, con procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando bigramas

Tabla 94 Experimentos con el conjunto de etiquetas completo, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.130 palabras, empleando bigramas

Tabla 95 Experimentos con el conjunto de etiquetas completo, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando bigramas

Tabla 96 Experimentos y gráfica de cobertura con el conjunto de etiquetas simplificado, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando trigramas

Tabla 97 Experimentos con el conjunto de etiquetas simplificado, con procesado especial de locuciones, sobre un conjunto de evaluación del dominio de discapacidad de 22.518 palabras, empleando bigramas

Tabla 98 Experimentos con el conjunto de etiquetas simplificado, con procesado especial de locuciones, sobre un conjunto de evaluación del dominio de discapacidad de 22.518 palabras, empleando trigramas

Tabla 99 Parámetros de predicción de F0 que presentan diferencias significativas al adoptar valor 1 frente a adoptar valor 0

Tabla 100 Modelado paramétrico de la curva de F0 para los nombres propios con una sola tónica (modelo de picos y valles)

Tabla 101 Modelado paramétrico de la curva de F0 para los nombres propios con varias tónicas (modelo de picos y valles)

Tabla 102 Modelado paramétrico de la curva de F0 para los nombres propios con varias tónicas (modelo de picos y valles)

Tabla 103 Modelado paramétrico de la curva de F0 para los sintagmas nominales en enunciativas con una tónica (modelo de picos y valles)

Capítulo 1 Introducción

La palabra tanto hablada como escrita ha supuesto para el hombre uno de sus mayores logros en la carrera hacia el dominio de la naturaleza. Ella le ha permitido desarrollar el más complejo mecanismo comunicativo existente en los seres vivos, capaz de los más altos niveles de abstracción. Tan poderosa herramienta, lejos de ser un fruto consciente de su racionalidad, ha sido la que ha cimentado (sino impulsado), el desarrollo de la misma. La naturaleza fundacional del habla (imbricada profundamente en los mecanismos del pensamiento) ha hecho muy difícil a lo largo de la historia tanto su estudio como su emulación.

Si definimos la síntesis de voz como la capacidad de los sistemas electrónicos para producir voz que parece humana, nos hallamos muy lejos de alcanzar tan ambicioso objetivo. La tecnología actual es capaz de convertir texto en voz con una alta tasa de inteligibilidad, aunque su grado de naturalidad no sea tan alto como desearíamos: no podemos imitar el amplio espectro de cadencias, melodías y cualidades que cubre la voz humana. Por lo general las voces sintéticas pueden ser catalogadas como monótonas o incluso aburridas: nuestros ordenadores carecen por ahora de capacidad para transmitirnos emociones, para adaptar la voz a diferentes estilos de locución (formales o informales), carecen de capacidad para engañarnos.

Vivimos una época en la que se está dando un gran auge en los estudios teórico-prácticos de la llamada inteligencia social o inteligencia emotiva, la que se encarga de controlar con inteligencia las propias emociones, reconocer las emociones de los demás y reaccionar empáticamente a las mismas. No es descabellado pensar en que esa misma inteligencia social debería gobernar las futuras aplicaciones de intercomunicación hombre-máquina, haciéndolas más y más amigables (*user-friendly*). Para ello, deberíamos de dotar a los sintetizadores de una voz más diversa y humana: un usuario habitual del sistema o un usuario con problemas que dialoga con el sistema, y tras repetidos intentos no consigue acceder a la información que precisa, deben ser tratado de un modo especial, como lo haría un experto humano. Adaptar la voz y el estilo al contexto en el que se emplea es uno de los principales objetivos que debemos alcanzar.

Como se señalaba hace ya muchos años en una conocida revista científica:

"Machines which, with more or less success, imitate human speech, are the most difficult to construct, so many are the agencies engaged in uttering even a single word -- so many are the inflections and variations of tone and articulation, that the mechanician finds his ingenuity taxed to the utmost to imitate them." (Scientific American, 14 January, 1871)

1.1 Objetivos de la Tesis

Dentro del campo de la conversión texto a voz, en esta Tesis nos proponemos abordar los siguientes objetivos

relacionados en general con la naturalidad y la variedad, todo ello aplicado al castellano:

1.1.1 Procesado lingüístico automático

§ Mejora del proceso de normalización de textos basándose en reglas manuales y diccionarios, buscando gran cobertura.

§ Mejora del sistema de categorización empleando 2 técnicas: etiquetado por reglas de transformación inferidas automáticamente a partir de un corpus etiquetado (técnica de Brill) y desambiguación por métodos estocásticos.

§ Creación y evaluación de una gramática para un análisis sintáctico de tipo *chunk*.

1.1.2 Modelado de la F0 para síntesis en dominio restringido

§ Diseño y creación de bases de datos aptas para el estudio de fenómenos segmentales y prosódicos en dominio restringido, mediante técnicas voraces.

§ Análisis y modelado de F0 de una base de datos de voz para dominio restringido, con estudio de los parámetros más relevantes, así como su modo de codificación.

1.1.3 Análisis y síntesis de habla con emociones

§ Diseño y grabación de una base de datos de voz emotiva simulada.

§ Modelado segmental y prosódico de una voz masculina parametrizable en castellano, empleando síntesis por formantes.

§ Modelado diferencial del habla emotiva a partir de habla neutra. Evaluación del habla con emociones en pruebas de identificación.

§ Análisis de la importancia de la cualidad de voz y de la prosodia en la identificación de voz simulando un estado emocional.

1.2 Contenido de la Tesis

La Tesis comienza exponiendo en el capítulo 2 el estado de la cuestión de la conversión texto habla en lo relativo a los distintos subtemas que la componen, desde el procesamiento lingüístico y prosódico hasta la conversión texto habla con emociones.

A continuación se analiza, en el capítulo 3, el primer problema de cualquier conversor, esto es, cómo normalizar el texto de entrada frente a la variación lingüística y de formato. A continuación se estudiarán diversas propuestas de etiquetadores morfosintácticos, con especial énfasis en la alternativa basada en modelado estocástico e información léxica genérica (en vez de ligada a un dominio concreto). Finalmente se expondrá y evaluará un analizador sintáctico de tipo sintagmático basado en reglas de experto y la estrategia de buscar el análisis más simple y compacto.

El capítulo 4 trata del modelado de F0 en un dominio restringido. Se comenzará desarrollando una nueva propuesta de diseño de una base de datos de voz orientada a prosodia a partir de los datos disponibles sobre el dominio, que incluya una importante moderación de su carácter voraz por medio del establecimiento de una secuencia de subobjetivos que permitan alcanzar el objetivo global de manera más precisa que una estrategia voraz pura. El resto del capítulo desarrolla la experimentación relativa al modelado de la curva de F0 mediante perceptrones multicapa, estudiando los parámetros más relevantes, especialmente las diferencias con modelados previos en dominios no restringidos.

En el capítulo 5 se estudiará una modalidad de conversión nueva en castellano, aunque con gran interés internacional en los últimos años: la síntesis de voz emotiva. Se expondrá el desarrollo de una nueva voz masculina personalizable empleando un sintetizador basado en formantes, que sirve de base para el posterior estudio del modelado prosódico y segmental de 3 emociones a partir de una base de datos de voz simulada especialmente diseñada y grabada en el marco de esta Tesis. Como paso previo hacia una futura síntesis emotiva por concatenación, se demostrará la necesidad de disponer de un modelado segmental muy preciso incluso en las

emociones dotadas de una prosodia característica.

Finalmente se expondrán las principales conclusiones así como las líneas futuras de trabajo que se abren como consecuencia de los trabajos de esta Tesis.

Capítulo 2 Estado de la cuestión

2.1 Introducción

Si las Tecnologías del Habla tienen por objetivo final investigar y desarrollar sistemas automáticos capaces de mantener diálogos orales con humanos, no cabe duda que la conversión texto a voz es una de sus áreas más importantes. Para hacer que el habla sintética sea aceptable para los humanos con los que interacciona, sus principales características deberían ser:

§ **Alta inteligibilidad**: los segmentos empleados deben ser claramente identificables por parte de oyentes humanos.

§ **Alta naturalidad**: alta calidad, esto es, alta similitud entre habla sintética y habla natural, lo cual debe incluir la capacidad para expresar estados de ánimo o intenciones. La variedad como opuesto a la monotonía, es una característica fundamental para dotar de humanidad a la voz sintética.

Los progresos en síntesis han conducido a sistemas con altas tasas de inteligibilidad y naturalidad creciente, especialmente a partir de la aparición de la técnicas PSOLA. La calidad alcanzada permite que la síntesis sea utilizada de manera total o parcial en multitud de aplicaciones:

§ **Máquinas de lectura para ciegos, lectores de correo electrónico o mensajes SMS**: en los cuales la inteligibilidad es primordial, aunque el paso del tiempo y la continuidad en el uso pueden hacer que sea la monotonía y la falta de naturalidad las que hagan inaceptables estas aplicaciones.

§ **Comunicadores para personas con discapacidad vocal**: que se convierten en sus prótesis vocales y que necesitarían la posibilidad de personalizar la voz y dotarla de la posibilidad de transmitir emociones (*J. M. Montero et al* 1998) *(G. Martínez-Salas* 1998). Es habitual que se vean acompañados de software de predicción de palabras o expresiones (*S. Palazuelos-Cagigas* 1994) (*S. Palazuelos-Cagigas* 2001).

§ **Sistemas vocales interactivos (IVR:** *Interactive Voice Response systems***)**: la mayoría de estos sistemas (mayoritariamente telefónicos) se basan en la reproducción de mensajes pregrabados de alta calidad y concatenan palabras o secuencias de palabras para obtener las expresiones deseadas con la prosodia adecuada: números, fechas, horas... Sin embargo, a la hora de leer conjuntos muy abiertos (como nombres propios, nombres de poblaciones, etc.) o conjuntos cuyos contenidos varían con frecuencia (como los mensajes de correo electrónico), la síntesis de voz se vuelve insustituible por razones económicas y de tiempo de actualización. *(A. Casas-Guijarro* 1997).

§ **Agentes dotados de vida artificial**: con el auge de las tecnologías de reconocimiento de voz y desarrollo de sistemas inteligentes, la síntesis de voz es un componente primordial para poder crear seres animados artificiales (robóticos o virtuales) con capacidad de comunicarse oralmente (*J. Cassell* 2000) (*J. Heras 2002*) (*J.M. Montero, M.M. Duque* 2003b).

El modo más sencillo de generar voz es reproducir mensajes humanos pregrabados. Este método proporciona alta calidad, y es el mejor y más utilizado cuando el vocabulario es bajo y estable (gran parte de las aplicaciones de IVR lo emplean). Es popular denominar voz sintética a la concatenación de mensajes (especialmente si no se ha

respetado la variedad prosódica), aunque en Tecnologías del Habla es más común usar el término voz sintética para voz generada sin restricciones cuya prosodia se adapta al contexto donde va a ser insertada.

La señal sintética, como la natural, puede analizarse según dos dimensiones acústicas:

§ **Segmental**: parámetros o características que afectan a cada segmento de voz particular, sea este un fonema, un difonema o un semifonema: formantes, *tilt*, etc. Tiene influencia tanto en la inteligibilidad como en la naturalidad.

§ **Suprasegmental**: parámetros de tono, temporales y dinámicos que no caracterizan intrínsecamente un segmento, sino que influyen sobre varios segmentos, estableciendo una curva prosódica. Estos parámetros son fundamentalmente F0, duración / ritmo e intensidad, aunque este último es mucho menos relevante a la hora de definir el acento en castellano (*E. Enríquez et al* 1989). Estos parámetros contribuyen más a mejorar o empeorar la naturalidad que la inteligibilidad (salvo casos extremos como una duración muy breve) y contribuyen a estructurar y organizar el discurso oral, ayudando a establecer relaciones y contrastes entre palabras o sintagmas.

2.1.1 Sistemas comerciales de conversión texto a voz

En esta Tesis haremos uso de 2 sistemas de conversión texto a voz en castellano:

§ **Infovox**: El sintetizador de *Telia* Infovox es un conocido sintetizador multilingüe disponible comercialmente en castellano (además de en sueco, inglés, etc.). Este sintetizador fue desarrollado por el *Royal Institute of Technology* de *KTH* como un sintetizador de formantes en cascada (*OVE*). La versión parcialmente desarrollada en esta Tesis (*Infovox 230*) posee la posibilidad de crear y guardar nuevas voces que permitan personalizar el habla para servir como comunicador.

§ **Boris**: es un sintetizador monolingüe por concatenación de difonemas y modificación prosódica en el dominio temporal, con modelo multiplicativo para el modelado de duraciones, modelo neuronal para el modelado de F0, y que disponía de un locutor en castellano masculino al comenzarse esta Tesis (*F. Giménez de los Galanes* 1995) (*J.M. Pardo et al* 1995).

2.2 Procesado lingüístico

2.2.1 Etiquetado morfosintáctico automático

El primer paso para la realización de un análisis gramatical de un texto es el etiquetado morfosintáctico (*POS tagging*) o categorización de sus palabras o expresiones *(S. Quazza & H. van der Heuvel* 2000). Podemos dividir este problema en 2 sub-problemas: las palabras fuera de vocabulario (*cobertura léxica*) y las palabras presentes en el diccionario (*desambiguación*).

2.2.1.1 Teoría lingüística generativa

A lo largo de esta tesis se emplearán reglas de origen lingüístico en diversas ocasiones y, aun cuando los objetivos no serán lingüísticos sino de ingeniería, es necesario reseñar brevemente la compleja historia de la teoría lingüística de *Chomsky* (teoría principalmente sintáctica), desde los años 50 hasta nuestros días.

De acuerdo con la jerarquía establecida por Chomsky (*N. Chomsky* 1959), las reglas de etiquetado que se emplearán en esta Tesis tienen el formato de las reglas propias de gramáticas tipo I, aunque el conjunto de reglas ordenadas que se empleará no constituyen una gramática descriptiva tipo I. Las gramáticas de reglas dependientes del contexto son centrales en la teoría fonológica de *Chomsky* (*N. Chomsky & M. Halle* 1968) y aparecerán en la síntesis de voz por formantes para generación de voz con emociones. Finalmente, las gramáticas de análisis sintagmático robusto que se emplearán, serán reglas de estructura de frase, independientes del contexto (gramática de tipo II).

Aún cuando las sucesivas versiones de la teoría han ido refutando buena parte de los análisis previos, muchos componentes permanecen aunque evolucionen profundamente:

§ **Gramática Generativa y Transformacional - Teoría Estándar**: emplea *reglas de estructura de frase (*de naturaleza descriptiva, no de procedimiento) y *transformaciones* sobre los árboles de

análisis sintáctico. Una crítica fundamental que realizar a este modelo (y que obligó a su posterior reformulación) es que la propia base matemática que la sustenta (y que constituye una importante formalización del procesamiento basado en reglas) revela que es poco probable desde el punto de vista de su aprendizaje por parte del cerebro, dado que su potencia es la de una máquina de *Turing*. Esta crítica dio lugar a la **Teoría Estándar Ampliada**, que reduce las transformaciones a movimientos de constituyentes (sintagmas) para limitar la sobrecapacidad del sistema, e introduce la *categoría vacía* y las *huellas*, que deben dejar los movimientos para permitir su reconstrucción.

§ **Teoría de Principios y Parámetros**: sólo permanece una transformación (*mover alfa*), limitada en su alcance por las denominadas *barreras*; las reglas dejan de ser el elemento central de la teoría, para dejar paso a los *principios* y *parámetros* que caracterizan la relación de cada gramática particular con la gramática universal (destaca el *principio de economía de la derivación:* se debe minimizar la cadena de estructuras y movimientos); el lexicón y su estructura de rasgos ganan importancia, recogiendo la información que gobierna los *papeles temáticos*, la *rección* y el *ligamiento* (entre huellas y antecedentes); la sintaxis *X con barra* sustituye a las reglas de estructura de frase, y establece la estructura general de los constituyentes: especificador (generalmente pre-nuclear en castellano), núcleo y complementos o adjuntos (preferentemente post-nucleares en castellano). Para nuestros propósitos de análisis sintáctico computacional, la sintaxis X con barra guiará parcialmente las reglas locales de estructura de constituyentes simples (a los que llamaremos segmentos).

§ **Programa Minimalista**: en (*N. Chomsky* 1994), aunque se trate simplemente de un programa de investigación (no de una teoría o un modelo gramatical), el proceso de principios se radicaliza aún más y busca una formulación mínima muy general y universal (el principio de economía es un principio de principios de primerísima importancia). Sólo existe un único nivel de representación fonológica superficial (aunque existe la representación lógica), con lo cual simplifica mucho el conjunto de principios. Los elementos de procesamiento del nuevo programa son *merge* y *agree* (que ocupan el lugar de las reglas) y *move* (copia de constituyentes que ocupa el lugar de las transformaciones y que no da lugar a huellas, que sólo se emplea como último recurso) que operan a partir de la información léxica y de rasgos disponible (operación *select*). Aunque en la presente Tesis emplearemos un principio de economía global relacionado con el Programa Minimalista, en ningún caso se tratará de una hipótesis fisiológico-realista, sino de una aproximación de naturaleza empírico-algorítmica (*D. Johnson & S. Lappin* 1997).

Algunas de las características de lenguas como el castellano han influido importantemente en esta evolución (*V. Demonte* 1991):

§ **Orden libre de los constituyentes**: las oraciones complejas y largas, con grandes posibilidades de variar su estructura superficial, han supuesto siempre un gran problema para el desarrollo de un conjunto robusto de reglas que analicen globalmente las oraciones.

§ **Sistema de casos léxicamente muy reducido**: en castellano apenas quedan palabras con marcas de caso (los pronombres) y la asignación de papeles temáticos se ha de basar en criterios de posición (o distancia al elemento rector), en las preposiciones y en la información de rección en el lexicón.

Aunque estos trabajos son de naturaleza fundamentalmente lingüística y sicológica[1], de estas ideas han surgido sistemas computacionales artificiales como *Generalized Phrase Structure Grammar*, *Head Phrase Structure Grammar* y *Lexical Functional Grammar*.

2.2.1.2 Preprocesamiento

El procesamiento de texto sin restricciones requiere la normalización del texto de entrada, procesando las palabras consideradas como no-estándar: nombres propios, fechas y horas, letras aisladas, palabras con signos de puntuación, acrónimos, expresiones numéricas (árabes o romanas), abreviaturas, nombres propios, palabras mal escritas, expresiones relacionadas con *Internet*, etc.

De acuerdo con (*A. Jiménez-Pozo* 1999), es necesario normalizar más de un 7,5 % de las palabras encontradas en un corpus periodístico de 4,5 millones de palabras, aunque en la mayoría de los casos se trata de nombres propios

de personas o entidades. Estas cifras, de todas maneras, son muy dependientes del dominio (*R. Sproat, et al* 1999).

2.2.1.3 Diccionarios y plataformas léxicas en castellano

En el proyecto *Poliglot* 2104 se desarrolló un sistema de bases de datos relacional para análisis morfológico que aplicó al castellano el GTH (UPM y UNED), aunque posee baja cobertura debido a su limitado vocabulario de lexemas.

En *CRATER* (*F. Sánchez-León & A. Nieto* 1997) el diccionario comprende más de 40.000 lemas que se expanden en más 440.000 formas (sin incluir enclíticos), aunque no ofrecen cifras sobre cobertura (*recall:* tanto por ciento de categorías correctamente predichas); coincide con el tanto por ciento de aciertos si sólo admitimos predecir una categoría por palabra.

El sistema *COES* es una adaptación al castellano del sistema léxico *ispell* (*S. Rodríguez & J. Carretero* 1996) que se basa en un diccionario de unas 80000 palabras y unas 3500 reglas morfológicas de conjugación y derivación obtenidas manualmente, con una cobertura superior al 97,5% y una precisión superior al 99,5%, trabajando a alta velocidad (1600 palabras por segundo).

La plataforma *ARIES* contiene un diccionario y un analizador morfológico propiamente dicho; con unos 60.000 lexemas que modelan más de 700.000 formas distintas (*Empresa Daedalus* 2001) (*J.M. Goñi, J.C. González y A. Moreno* 1997) (*A. Moreno & J.M. Goñi* 1995) (*A. Martín & J.M. Goñi* 1995) (*J.M. Goñi* 1998).

Tras el proyecto *ITEM*, el diccionario del *Centre de Llenguatge i Computació* contiene ya más de 107.000 lemas (15000 verbales, 89000 nominales o adjetivales y 3000 ejemplos de categorías cerradas y locuciones) y más de un 1.000.000 formas (diccionario indexado con *caché*), ofreciendo una cobertura del 99,5 por ciento con una tasa de 1,64 categorías por palabra (*J. Atserias et al* 1998). Los modelos morfológicos para la generación del diccionario (*MACO+*) incluyen 29 paradigmas nominales o adjetivales y 45 verbales. El pre-procesamiento incluye fechas, abreviaturas, nombres propios, expresiones numéricas y locuciones (*J. Carmona et al* 1998).

El *Sistema de diccionarios electrónicos del español* (*C. Subirats* 1998) es un sistema léxico con más de 550.000 formas simples y unas 50.000 locuciones o formas léxicas multipalabra, destacando el grafo ambiguo de representación de una frase con locuciones complejas (http://elies.rediris.es/elies10/1.htm).

En el GEDLC han desarrollado un sistema morfoléxico ambicioso basado en la mayoría de los diccionarios electrónicos existentes en castellano, y con un completo conjugador verbal (*O. Santana et al* 2001), que incluye ya cierta información sobre la rección de más de 4.000 verbos.

El sistema GALENA (*J. Graña* 2000) compila el diccionario en forma de autómata mínimo de forma similar a (J. *Hopcroft & J. Ullman* 1979) (*J. M. Montero* 1992), aunque su cobertura parece ser bastante menor que los sistemas anteriores.

La única prueba de cobertura léxica entre estos sistemas que el autor conoce aparece en (*J.C. González, J.M. Goñi & A. Nieto* 1995), donde ARIES consigue un 0,6 % de palabras desconocidas (frente al 1,4% de COES), aunque en la evaluación se incluye un pequeño diccionario especializado en el dominio de evaluación, y se excluyen nombres propios y palabras mal escritas[2].

La formación de palabras en castellano es un fenómeno extraordinariamente productivo y complejo (composición y morfología derivativa: prefijación, sufijación apreciativa y no apreciativa, derivación regresiva), debido a su irregularidad y a las variaciones alofónicas, y no suele ser tratada de manera completa en los sistemas (*J. Vilares, D. Cabrero & M. A. Alonso* 2001).

Este recorrido por los sistemas disponibles parece señalar que estrategias léxica más complejas como la morfología en 2 niveles (superficial y teórico) que emplea reglas de trasformación (*K. Koskenniemi* 1983), debido a su complejidad, no compensan a la hora de identificar las palabras fuera de vocabulario, dado que estas palabras acostumbran a tener una morfología más regular.

2.2.1.4 Técnicas de desambiguación en el etiquetado morfosintáctico

Existen diversas escuelas dentro del procesamiento de lenguaje natural, no siempre disjuntas: métodos estocásticos, métodos simbólicos, clasificadores, métodos automáticos basados en corpus (*data driven*), métodos

manuales:

1) **Métodos estocásticos basados en corpus**:

- **Basados en modelos de Markov (MM)** discretos de orden 1 o 2 y en frecuencias de coocurrencia (*B. Merialdo* 1994*)* (*T. Brants* 2000); son equivalentes a los llamados n-gramas con programación dinámica. Como suele suceder con las técnicas estocásticas suele ser necesario suavizar los modelos para tener en cuenta la escasez de los datos (*data sparsity*), produciéndose significativos cambios según el tipo de suavizado, como se ejemplifica en (*J. Zavrel & W. Daelemans* 2001). También es habitual el empleo de heurísticos o *biases*, para mejorar su tasa máxima, produciéndose así una convergencia con métodos basados en conocimiento. En *(T. Brants* 1999) se describe un sistema de gran calidad (con una tasa en torno al 3 % en Lancaster Oslo Bergen corpus o en Wall Street Journal corpus), que emplea trigramas, suavizado por interpolación lineal (*deleted interpolation*) y búsqueda limitada por un haz (*beam search)*. De acuerdo con las tasas de aprendizaje, un corpus de menos de 200.000 palabras es suficiente para la estimación del modelo de trigramas, especialmente si se dispusiese de un diccionario adaptado al dominio y de gran cobertura que minimice el número de palabras fuera de vocabulario (*OOV*: *Out Of Vocabulary words)*. Una de las ventajas de este método estocástico es que permite ordenar hipótesis, pudiendo ser utilizado en procesos de desambiguación no completa (sobre WSJ se puede así conseguir una cobertura superior al 99 %, reduciendo la ambigüedad remanente del 55% al 12%). También nos permiten obtener un ordenamiento de las posibles soluciones, así como obtener medidas de confianza sobre el etiquetado obtenido. Aunque por medio de la programación dinámica es capaz de modelar fenómenos globales como la propagación de género y número, no puede modelar otras relaciones de larga distancia como la rección verbal, aunque se trata de técnicas inevitables por su robustez.

- **Basados en el principio de la máxima entropía (ME)**: que postula que el mejor estimador es aquel que otorga máxima entropía a los datos no vistos, sujeto a la restricciones impuestas por los datos vistos (típicamente en una ventana de 5 palabras). Con ello combina un buen modelado contextual local con un modelado probabilístico. El sistema *JMX* alcanza una tasa en torno al 3 %, como los mejores sistemas disponibles en inglés (*Ratnaparkhi* 1996). Según algunos experimentos, es posible que, para conseguir buenos resultados, necesita un corpus mayor que los métodos basados en *Markov* (*B. Megyesi* 2001).

2) **Clasificadores basados en corpus:**

- **Árboles de decisión y redes lineales o neuronales:** estos conocidos clasificadores han sido aplicados al etiquetado, aunque en menor medida que otras técnicas (*L. Padró* 1997) (*H Schmid* 1994). El algoritmo C 5.0 alcanzó un 97,97 % en el corpus *LOB* (*W. Daelemans, A. van den Bosch & J. Zavrel* 1999). Puede entrenarse un único clasificador para todos los casos o varios clasificadores especializados en cada tipo de ambigüedad, se puede emplear el contexto izquierdo y el derecho, o sólo el izquierdo (el ya desambiguado). Frente a los clasificadores estocásticos basados en programación dinámica, presentan el defecto de emplear únicamente información local y que nunca buscan la secuencia que encuentra la mejor secuencia global de soluciones locales.

- **Métodos basados en memoria**: en lugar de intentar abstraer conocimiento a partir de un corpus, estas técnicas se basan en la acumulación de ejemplos correctamente clasificados sin resumirlos o generalizarlos (sin perder excepciones), y en una distancia basada en la similitud de los ejemplos y en la ganancia de información que aporta cada rasgo (*W. Daelemans & J. Zavrel* 1996). A la hora de hacer la clasificación de nuevos ejemplos se usa la técnica de los vecinos más próximos (*k-Nearest Neighbour*) y los costes son previamente estimados a partir del corpus de entrenamiento (*R. San Segundo, J.M. Montero et al* 2000). Los rasgos empleados incluyen las 2 etiquetas ya desambiguadas (de las 2 palabras previas), la palabra actual (mejor que las posibles etiquetas de la palabra actual) y de la palabra o 2 palabras siguientes, alcanzando un 97,86 % sobre el corpus LOB (*W. Daelemans et al* 1999).

3) **Métodos simbólicos basados en reglas** *if ... then ...*

- **Reglas de transformación**: inicialmente se le asigna a cada palabra su etiqueta más probable según un diccionario; luego se va transformando esta etiqueta en función del contexto. Las reglas de transformación poseen más potencia descriptiva que los árboles de decisión (a los que incluyen), aunque las técnicas de extracción automática pueden hacer que los segundos superen a las primeras (*D. Palmer, J. Burger & M. Ostendorf* 1999). Los rasgos que se verifican en este método TBL (*Transformation Based Learning)* son: las etiquetas de las palabras circundantes (hasta 3), así como las palabras mismas. Aunque la técnica básica asigna una única etiqueta por palabra, se extendió la técnica para añadir etiquetas al etiquetado básico, alcanzando una cobertura del 98,4 % con una ambigüedad remanente del 19% (*E. Brill* 1995). Las principales objeciones que realizar a la técnica *TBL* son el coste computacional de su entrenamiento (crece cuadráticamente con el número de palabras multiplicado por el número de posibles patrones de reglas) y el hecho de que, partiendo de una información morfosintáctica completa en forma de lexicón, se quede con la etiqueta más probable de cada palabra, procediendo a continuación a corregir algunos de los errores cometidos; las reglas de transformación no tienen en cuenta las posibilidades incluidas en el diccionario. Las reglas de transformación pueden combinarse con otros tipos de etiquetadores, dado que pueden aprender cómo mejorar un etiquetado (*Q. Ma et al* 1999), aunque posiblemente deben incorporarse patrones de reglas que compensen la información empleada por el etiquetador que sirve de base (por ejemplo, reglas de largo alcance para el tratamiento de la rección, para complementar la información local de un modelo de trigramas). Para lenguajes con un fuerte componente morfológico como el castellano, puede ser necesario emplear reglas de naturaleza morfológica, reglas en las que cambiar un sufijo por otro debe dar lugar a una palabra presente en el diccionario, reglas que permitan modelar la conjugación sin necesidad de un lexicón con cientos de miles de formas, de manera similar a lo que se describe en (*A. Mikheev* 1997).

- **Reglas de restricción**: inicialmente se asignan a cada palabra todas las etiquetas que puede tener y, posteriormente, las reglas eliminan las que son incompatibles con el contexto. La obtención de reglas por técnicas de aprendizaje automático es especialmente dificultosa dado que se debe aprender qué fenómenos no se dan nunca, lo cual se suele ver seriamente afectado por la escasez de datos (*sparsity*). Así en (*N. Lindberg & M. Eineborg* 1998) emplean el sistema CProgol 4.2, con ventanas de 7 palabras, para alcanzar una cobertura en torno al 97 %, pero manteniendo un 15 % de ambigüedad remanente. Si se emplea la información del lexicón se puede superar el 96 % de precisión (*J. Cussens* 1997). Es importante limitar la creación de reglas mediante 2 parámetros: el número mínimo de ejemplos de aprendizaje por regla (número de ejemplos de entrenamiento) y la confianza en cada regla (número de aciertos / número de ejemplos de entrenamiento).

- **Reglas de selección**: inicialmente se asignan a cada palabra todas las etiquetas que pueda tener y, posteriormente, las reglas seleccionan las más adecuadas según el contexto. Son más fáciles de aprender porque se basan en fenómenos positivos (*L. Dehaspe & L. De Raedt* 1997), siempre que no se extienda mucho el contexto de aplicación (como sucede con los n-gramas). Suelen ser menos relevantes lingüísticamente que las reglas de restricción. Sería interesante que incorporasen información sobre probabilidades de aplicación que permitiesen una optimización global.

Estos tipos de reglas simbólicas se pueden extraer:

§ **Automáticamente**: empleando una estrategia basada en buscar en cada momento la regla de transformación que más reduce la tasa de error como en (*E. Brill* 1995) (*M. Hepple* 2000), o una estrategia *Monte Carlo* de muestreo del espacio de reglas posibles (*K. Samuel* 1998).

§ **Manualmente**: en (*J.P. Chanod et al* 1995) se cuestiona la mayor velocidad de desarrollo de los métodos estocásticos, aunque en el trabajo intervienen autores con fuerte experiencia en procesadores lingüísticos.

§ **Con una mezcla de reglas manuales y automáticas**: siempre costosa (*G. Schneider & M. Volk* 1998).

Existen 2 posibles estrategias de aplicación de reglas:

§ **Recorrido primero de reglas y luego de palabras**: en régimen mono-paso *(Rogelio Vargas 1996*) o en régimen multipaso *(A. Jiménez Pozo* 1999). Son especialmente interesantes las estrategias multipaso porque tienden a ser más descriptivas y con menos interacciones.

§ **Recorrido inverso de palabras y reglas**: (*J. Vergne* 2000).

Los mejores resultados obtenidos internacionalmente (*C. Samuelsson & A. Voutilainen* 1997) corresponden al sistema *ENGCG* (*English Constraint Grammar*) con más de 1000 reglas restrictivas y una cobertura superior al 99,5 % (aunque con una ambigüedad remanente del 8 %). Añadiendo reglas de carácter heurístico (entre 200 en ENGCG y unas 4000 en *ENGCG-2*), se reduce la ambigüedad a la mitad, pero a costa de duplicar el error (*P. Tapanainen & A. Voutilainen* 1994). De todas formas, estos resultados significativamente superiores a los de los métodos automáticos deben ser considerados con prudencia debido a:

§ Emplear un *tagset* diferente a los estándar (como el *Penn Treebank* de 36+12 etiquetas empleado en WSJ, o el *Brown* extendido de 135 empleado en LOB*),* con importantes diferencias en el tratamiento las formas *–ing* y *-ed*, la ambigüedad sustantivo / adjetivo o la ambigüedad entre nombres propios / comunes / acrónimos.

§ Emplear medidas de comparación de etiquetas que reducen posibles diferencias: no se distingue entre preposiciones y conjunciones.

§ Emplear un corpus de evaluación muy homogéneo, revisado por varios expertos humanos.

2.2.1.5 Sistemas combinados o integrados

En los últimos años se ha conseguido mejorar la tasa de etiquetado morfosintáctico combinando varios etiquetadores que empleen técnicas diferentes (estocásticos, de reglas de transformación, de máxima entropía, basados en memoria...) y que, por los resultados, resultan complementarios cuando se entrenan con los mismos datos (*E. Brill & J. Wu* 1998). En (*H. van Halteren et al.* 2001) se mejora y profundiza en las técnicas empleadas en (*H. van Halteren et al* 1998) y se entrenan clasificadores para combinar la salida de los etiquetadores, consiguiendo reducir la tasa de error casi un 25 por ciento respecto al mejor de los etiquetadores aislado, aunque ya se consiguen mejoras empleando simplemente decisión por mayoría entre los etiquetadores de base.

Especialmente interesante resulta la posibilidad de integrar diferentes fuentes de conocimiento en un único sistema: reglas manuales, árboles de decisión, probabilidades, etc. En (*L. Padró* 1997) se emplean técnicas de etiquetado por relajación entrenadas mediante un algoritmo iterativo de tipo *hill-climbing*, mediante las cuales se puede emplear cualquier conocimiento que se pueda expresar en forma de restricciones (*constraints*): bigramas, trigramas, reglas manuales, etc. Sobre WSJ supera el 97 % de acierto (*L. Márquez & L. Padró* 1997).

El uso de múltiples clasificadores puede ser igualmente empleado para detectar errores en la etiquetación manual (*H. Berthelsen & B. Megyesi* 2000).

2.2.1.6 Medidas de evaluación y comparación entre sistemas

Las principales medidas sobre la calidad de un sistema de etiquetado se basan en comparar el etiquetado obtenido por un sistema con un etiquetado manual independiente que sirve como referencia, donde sólo el corpus de referencia ha de ser obligatoriamente etiquetado de forma no ambigua. Destacan:

§ *Cobertura (recall)*: es el número de etiquetas correctas dividido por el número total de elementos etiquetados.

§ *Precisión*: se calcula como el número de etiquetas correctas dividido por el número total de etiquetas asignadas

§ *Factor F*: combina las medidas anteriores mediante la fórmula $(B^2+1)*P*R/(B^2*P+R)$, que, si le damos igual peso a ambas medidas (B=1), queda reducida a $2*P*R/(P+R)$.

La comparación entre sistemas de etiquetado morfosintáctico se ve dificultada por la necesidad de igualdad en el idioma, el dominio de trabajo y el conjunto de etiquetas gramaticales empleadas. Así para el idioma inglés se ha definido un corpus que sirve de marco de comparación entre sistemas (*Brown Corpus*) con 87 etiquetas, pero no existe su equivalente en castellano, aunque la parte del corpus *CREA* etiquetada manualmente podría realizar esta labor (*F. Sánchez-León et al* 1999).

En el sintetizador *Boris* (*J.M. Pardo et al*. 1995) se emplea un conjunto reducido de 38 etiquetas[3] que apenas distingue entre diferentes formas verbales, y no incluye rasgos tales como género / número / persona, caso, tiempo, modo, etc. Este conjunto es similar al empleado en el sintetizador *AMIGO* (*M. Rodríguez et al* 1994). En el proyecto ESPRIT 860 se definió el conjunto más detallado de 400 etiquetas para el etiquetado morfosintáctico, que es el que se usará en esta Tesis (*J. Pastor et al.* 1998).

Sin embargo, el empleo de un conjunto teóricamente elevado de etiquetas puede verse bastante reducido en la práctica, conduciendo a que muchas etiquetas pueden ser excluidas por ser muy poco frecuentes en un determinado dominio de trabajo (*P. Tapanainen & A. Voutilainen* 1994).

Aunque un conjunto más detallado de etiquetas podría parecer que dificulta el etiquetado morfosintáctico, esto no es incondicionalmente cierto[4], especialmente si:

- **se emplean etiquetas lexicalizadas**: muchos de los diferentes tipos de conjunciones se pueden distinguir unos de otros empleando listas; y lo mismo puede decirse de los distintos signos de puntuación, de las distintas formas de abreviaturas, de los verbos auxiliares o modales, etc.
- **se emplean etiquetas de palabra única**: si empleamos una etiqueta específica para la palabra *que*, para la palabra *uno*, o para la palabra *como*, conseguimos reducir la dificultad del problema de etiquetado automático. Otras veces la etiqueta de palabra única añade una etiqueta sin añadir complejidad (el pronombre *se*, la partícula *no*, el interrogativo *cómo*, las contracciones, la preposición *sin*,..)

El proyecto *EAGLES* (*G. Leech et al 1*998) define un conjunto de etiquetas y rasgos que pretende recoger la posible diversidad morfosintáctica de diversos idiomas, y que sirve de referencia para los conjuntos de etiquetas:

- **Nominales**: de varios tipos (propios, comunes) con género, número y caso (esto último no afecta al castellano).
- **Verbales**: con persona, género, número, formas no personales (*finite*), modo, tiempo, voz y estatus.
- **Determinantes / pronombres**: con género, número, persona, poseedor, caso (en castellano sólo queda un residuo del sistema de casos en los pronombres), tipo de pronombre (demostrativo, posesivo, personal...), tipo de determinante (demostrativo, posesivo...).
- **Adjetivales**: con grado, género, número y caso (no afecta).
- **Adverbiales**: con grado.
- **Artículos**: con tipo (determinado o indeterminado), género, número y caso (no afecta).
- **Preposiciones**: con tipo.
- **Numerales**: con tipo (cardinal u ordinal), género, número, caso (no aplica) y función.
- **Conjuntivas**: con tipo.
- **Interjecciones**.
- **Categorías únicas**: partícula negativa, marca de infinitivo (no afecta), etc.
- **Residuales**: tales como signos de puntuación, palabras extranjeras, etc.

Como reglas generales se recomienda disponer de etiquetas para tratar las locuciones (*ditto tags*) y también se recomienda etiquetar la abreviaturas y acrónimos como lo que son cuando se expanden. Es imprescindible contar

con un completo manual de normas y ejemplos que nos ayuden a garantizar la homogeneidad de criterios entre las diversas personas que etiqueten.

En el proyecto *CRATER* se emplea un conjunto detallado de etiquetas (más de 400). En el proyecto *MULTEXT* se inspiran en *CRATER* y siguen el modelo de *EAGLES*. Podemos destacar: la posibilidad de enclíticos en los verbos, la distinción entre 5 tipos de verbos (*ser, estar, haber, modales y principales*), los posesivos forman parte de los adjetivos, los pronombres y los determinantes, estos presentan 4 posibles casos (*nominativo, dativo, acusativo y oblicuo – conmigo-*), hay 4 tipos de determinantes (*demostrativos, indefinidos, posesivos e interrogativos*), hay 2 tipos de preposiciones (*simples o compuestas*), hay 2 tipos de adverbios (*general y la partícula no*), y 2 tipos de conjunciones (*coordinadas y subordinadas*).

En el proyecto *ITEM* (*J. Atserias et al* 1998) se emplea un conjunto de etiquetas amplio basado en el proyecto *PAROLE LE2-4017*, que sigue las directrices de *EAGLES*.

2.2.1.7 Etiquetado manual

Como no podía ser de otra manera, los *corpora* etiquetados manualmente contienen errores de etiquetado morfosintáctico. En experimentos de etiquetado por parte de varias personas en paralelo, no se alcanzan un acuerdo superior al 99,3 por ciento (*A. Voutilainen & T. Järvinen* 1995). Sobre las inconsistencias que presenta un corpus especialmente ruidoso como *Wall Street Journal* en (*H. van Halteren et al* 2001) se destacan: errores en asignación de número a un nombre propio o a un nombre común, el etiquetado morfosintáctico de *about, ago* o *more*, la clasificación de los participios como tales o como adjetivos o la etiquetación de complementos nominales pre-nucleares.

Una de las labores que describirá la presente Tesis será detectar errores e inconsistencias en el corpus 860 que emplearemos para nuestros experimentos.

2.2.1.8 Corpora en castellano

Los corpora disponibles en castellano son menores que los disponibles en inglés, como *Penn Treebank Wall Street Journal* o *Lancaster Oslo Bergen* (más de 1 millón de palabras).

El corpus del Proyecto *Esprit* I 291/860 (*UN-CS0588*: *"Linguistic analysis of European languages"*) está compuesto por 3 tipos de textos: textos de periódicos, legislación y legislación europea. Cuenta con más de 300.000 palabras y emplea un conjunto de más de 400 etiquetas basadas en 2 niveles y en rasgos (*J. Pastor et al* 1998). El primer nivel contiene 9 categorías primarias (verbo, nombre, adjetivo, adverbio, pronombre, preposición, conjunción, artículo y miscelánea), cada una de las cuales da lugar a 67 etiquetas secundarias (verbo transitivo, nombre propio, adjetivo posesivo, pronombre indefinido, preposición con artículo, artículo indefinido, conjunción copulativa, miscelánea abreviatura, etc.). Los rasgos contemplados son tiempo, modo, persona, y enclíticos verbales, género y número, grado de adjetivos y adverbios y tipo de locución. En (*J. Pastor* 1998) se expone una agrupación de las etiquetas en 160 macro-categorías para la obtención de mejores modelos de lenguaje.

El corpus del proyecto *CRATER*: contiene unas 500.000 palabras etiquetadas semiautomáticamente procedentes de la *ITU* (Unión Internacional de las Telecomunicaciones), con un conjunto detallado de etiquetas (más de 400).

Los proyectos *LEXESP: Base de datos informatizada de la lengua española* (*J. Carmona et al* 1998) emplea más de 230 etiquetas (62 categorías o subcategorías) y contendrá 5,5 millones de palabras, aunque sólo un sub-corpus de 84.000 palabras han sido revisadas a mano. El objetivo actual es disponer de un corpus con un millón de palabras desambiguadas manualmente.

El proyecto *CREA-CORDE* de la RAE ha sido financiado por CICyT desde 1996 y su investigador principal es Guillermo Rojo (*F. Sánchez León* 1998) (*M. Municio et al* 2000). En él se proyecta etiquetar manualmente un millón de palabras, con un diccionario que asigna un elevado número medio de etiquetas por palabra (1,74). Los textos comprenden los medios periodístico, literario, radiofónico, etc. Emplean el etiquetador por reglas manuales *RTAG*.

2.2.1.9 Sistemas de desambiguación en castellano

SPOST: forma parte de la arquitectura del sistema *Panglizer* y emplea 75 reglas contextuales de transformación

(no reglas de selección), obtenidas manualmente, ordenadas y aplicadas de manera multipaso; léxicamente cuenta con un diccionario con 65.000 lemas (basado en el diccionario *Collins*), con 2.000 locuciones y morfología verbal por concatenación (sobre-genera formas verbales). En un experimento reducido, sobre algo más de 10.000 palabras, alcanza el 94,41 por ciento en etiquetado morfosintáctico básico (POS) y sólo un 89,89 al incluir morfología. Destacan las confusiones entre adjetivos y sustantivos, nombres propios y comunes, y también entre artículos y pronombres. Esto último se explica porque consideran como pronombres los artículos que encabezan sintagmas determinantes como "el que" (*D. Farwell* 2001).

CRATER (*F. Sánchez-León & A. Nieto* 1997) (*F. Sánchez León* 1997): con un conjunto de 475 etiquetas y hasta 903666 palabras etiquetadas para entrenar, alcanza un 96 por ciento de precisión (sin tener en cuenta palabras extranjeras), destacando la aportación de los heurísticos que contribuyen a mejorar la tasa alcanzada por el modelo probabilístico, aunque el conjunto de *test* es muy reducido (algo más de 9.300 palabras). El sistema de sufijos para el etiquetado morfosintáctico de palabras desconocidas no es un objetivo importante dentro del proyecto, por lo que es evaluado con sólo 27 palabras. Los enclíticos se tratan por medio de un conjunto de reglas que recogen las excepciones de concatenación de la primera y segunda personas del imperativo. Incluye un módulo de pre-procesamiento de locuciones (*Multi Word Units*).

En (*E. Dermatas & G. Kokkinakis* 1995) se aplica un sistema probabilístico al castellano y a otros 6 idiomas europeos y, con el corpus 860, obtiene una tasa de error inferior al 4 % sobre las 11 etiquetas de primer nivel, e inferior al 6 % sobre el conjunto completo de etiquetas. Las tasas de error para palabras fuera de vocabulario son menores que el 30 % y el 45 % respectivamente.

En (*J. Zavrel & W. Daelemans* 1999) se aplican técnicas MBL sobre el corpus *CRATER* (800.000 palabras, aunque el etiquetado es en parte manual y en parte automático, con 484 etiquetas), obteniendo una tasa de acierto del 97,8 %. Los 6 rasgos empleados son: las etiquetas desambiguadas de las 2 palabras a la izquierda, la palabra y la etiqueta ambigua (*ambiguity class*) de la palabra que desambiguar y las etiquetas ambiguas de las 2 palabras a la derecha[5].

SpaCG-2 emplea una gramática de restricciones, sólo 14 etiquetas y, por los ejemplos disponibles, no distingue entre nombre propios y comunes, ni entre determinantes (o pronombres) demostrativos o posesivos, no emplea locuciones, etiqueta de un modo muy discutible estructuras como *el que o cada una* (pronombre + pronombre) o *dos mil* (numeral + sustantivo) o comete errores triviales como etiquetar la palabra *tercer* como infinitivo.

Sistemas en torno a *LEXESP*: en (*F. Pla & N. Prieto* 1998) alcanzan una tasa de error del 3 % empleando modelos de bigramas (*SLM toolkit*) o árboles de decisión (*L. Márquez* 1999), y un error del 3,36 % empleando inferencia de autómatas mediante ECGI. Al incluir información léxica en los bigramas (45 palabras muy frecuentes), se reduce el error hasta 2,58 % (*F. Pla, A. Molina & N. Prieto* 2001). Cifra similar se obtiene al emplear técnica de relajación para combinar bigramas y trigramas.

El sistema *GALENA* emplea modelos de Markov de orden 2 con diversos métodos de suavizado (*back-off* e interpolado lineal) y permite añadir diccionarios externos al corpus de entrenamiento. En (*J. Graña* 2000) procesa el corpus *CRATER*, aunque lo re-etiqueta con un juego de 373 etiquetas en el estilo *EAGLES*. Se comparan diversos etiquetadores sobre dicho corpus (*TBL*, *JMX*, *TnT* y el propio Galena). Si se usa el diccionario del corpus de entrenamiento, todos los etiquetadores alcanzan una precisión en torno al 98 %, con una cobertura superior al 96 % (el mejor resultado lo alcanza *JMX*: 96,676 % y el peor *Brill*: 96,211%, empleando 32.400 palabras para entrenar y 162 para evaluar). Si se emplea el sistema léxico adicional *Galena*, el propio sistema *Galena* alcanza 98,067 % - 96,693 %; si se emplea *TBL* con el mismo diccionario se consigue 98,027% - 96,618%. Un diccionario completo (entrenamiento + evaluación), esto es, sin palabras fuera de vocabulario, permite que *Galena* obtenga 98,681 % - 97,198 % y *TBL* obtenga 98,465 % – 96,738 %. Como se ve, se puede conseguir una mejora del 25 % en reducción de error, si no hay palabras fuera de vocabulario[6].

También en (J. Graña 2000) se emplea análisis probabilístico CYK extendido con reglas anulables para desambiguar textos en inglés. La técnica ascendente basada en *chart* permite implementar sencillos mecanismos de robustez, buscando las secuencias más largas y más probables, aunque los resultados sólo son buenos en un corpus donde la gramática empleada presenta una alta cobertura (no se realizan experimentos sobre un corpus de evaluación).

2.2.2 Sintaxis y análisis sintagmático

La importancia de la sintaxis y el procesamiento lingüístico en la síntesis de voz se ve reflejada en que casi cualquier sistema dispone de algún tipo de módulo lingüístico (*D. Klatt* 1987) (*J.M. Pardo et al* 1995). Sin embargo, las relaciones entre sintaxis y prosodia son complejas, aunque una estructura sintáctica plana parece tener mayor relación con la estructura prosódica que una estructura más jerárquica y compleja, con dependencias funcionales (*S. Abney* 1996) (*E. Selkirk* 1994) (*J. Bachenko & E. Fitzpatrick* 1990). En general, las duraciones, el pausado y la entonación se ven influidos por agrupaciones textuales que tienden a no tener en cuenta la estructura sintáctico-funcional completa, sino las agrupaciones de palabras-función en torno a las palabras-contenido (*J.P. Gee & F. Grosjean* 1983).

Los segmentos prosódicos se caracterizan por la presencia de un acento principal (palabra contenido) y varios posibles acentos secundarios o palabras desacentuadas (palabras función), y por delimitar posibles grupos fónicos (cada grupo fónico contendrían uno o varios segmentos prosódicos). La estructura interna de estos segmentos se puede definir cómodamente con una gramática de contexto libre[7] con concordancias *(D. Polanco* 2000).

Entre los nombres que ha recibido este tipo de procesamiento sintagmático (o de sintaxis plana) podemos destacar: *partial parsing, clause-bracketing, shallow parsing, chunk parsing, surface-oriented parsing*. Y los fragmentos de texto son llamados: *chunks, non-recursive clauses, non-recursive phrases, core phrases, maximal-length noun phrases, terminological units*. Ejemplos de segmentos son el nominal, el verbal, el de infinitivo, el adjetival, el de gerundio o el de adverbio.

Este tipo de procesamiento de orientación prosódica y de síntesis tiene su equivalente dentro del campo del reconocimiento y comprensión de lenguaje natural en los sistemas basados en conceptos, en los cuales se segmenta cada petición en una secuencia de conceptos no recursivos tales como origen, destino, fecha, tipo de billete, etc. que admiten un análisis sintáctico (*parsing)* robusto, una sintaxis interna de estados finitos (incluso probabilística) y una sintaxis global en el nivel de concepto (equivalente *grosso modo* al de segmento). Esta combinación de estrategias permite incluso responder preguntas compuestas por coordinación o subordinación, de cierta complejidad como en (*J. Colás* 1999) donde se emplea un analizador de contexto libre para procesar el nivel de conceptos y relaciones entre ellos.

2.2.2.1 Características de los segmentos o sintagmas simples

§ **No recursivos**: un segmento nominal no puede incluir otros segmentos nominales directamente, o a través de un complemento preposicional.

§ **No solapados y máximos**: un segmento no puede estar incluido en otro; un adjetivo modificando a un sustantivo (dentro de un segmento nominal) no puede constituir un segmento.

§ **Estructura nuclear interna**: los segmentos verbales tienen, al menos, un verbo como núcleo.

§ **Relaciones externas no determinadas**: para este tipo de análisis los segmentos son independientes entre sí.

§ **Posibilidad de análisis sintáctico de complejidad lineal basado en decisiones locales**: aunque resulte elegante describir los segmentos mediante gramáticas de tipo II (*Context-Free Grammar*), es posible realizar la segmentación linealmente o casi linealmente con la longitud de la oración.

§ **Robustez**: la gran mayoría de las oraciones son procesadas y la mayoría de los segmentos son detectados.

En las definiciones iniciales de *Abney* se excluyen los complementos post-nucleares y hay ciertas partículas que no pertenecen a ningún segmento (preposiciones, conjunciones, signos de puntuación), aunque hay excepciones (coordinación pre-nominal) poco explicadas. Los pronombres personales sujeto y los adjetivos (como atributos) o los adverbios adjuntos forman parte del segmento verbal a cuyo núcleo acompañan, aunque son posiciones donde se puede introducir, enfáticamente, una pausa. El genitivo sajón constituye un final de segmento a pesar de su clara ligazón con el núcleo que lo sucede (*S. Abney* 1996).

Se trata de un análisis sintáctico de superficie (*surface-based*) que segmenta la oración en islas locales no

ambiguas, que puede servir de base para un proceso posterior de *attachment* (*S. Abney* 1996). Los principios comúnmente utilizados son: el de economía derivativa, el de mínima longitud de descripción (*Minimun Description Length*, optimizando globalmente) o *longest-matching rule* (optimizando localmente). Según este principio el mejor análisis, en la inmensa mayoría de los casos, es aquel que resulta en una descripción más compacta y sencilla, más elegante (*J. Goldsmith* 2001).

2.2.2.2 Sistemas automáticos de segmentación en sintagmas simples

Las técnicas que se han empleado son bastante similares a las empleadas en el etiquetado morfosintáctico gramatical:

- **Técnicas estocásticas**: (*K. Church* 1988)
- **Técnicas basadas en reglas:** contextuales (*J. Vergne* 2000) o basadas en autómatas o transductores finitos (*J. P. Chanod & P. Tapanainen* 1996)
- **Técnicas conexionistas**: (*C. Lyon & B. Dickerson* 1995)

A pesar de tratarse de un concepto bastante intuitivo (*S. Abney* 1997), presente de una u otra manera en todas las teorías y modelos lingüísticos (el concepto de constituyente oracional simple), los detalles dan lugar a interpretaciones muy diversas; así un importante problema al comparar sistemas reside en la diversidad de definiciones implícitas de un segmento que cada autor hace. En (*A. Voutilanen* 1993) se emplean reglas de restricción manuales y un diccionario que relaciona cada palabra con sus posibles posiciones dentro de uno o varios segmentos, y donde las reglas explicitan aquellos contextos que no son factibles para determinar dónde no puede comenzar y acabar un segmento y por tanto conseguir la segmentación; de los ejemplos de segmentación que expone como correctos se deduce que excluye algunos complementos pre-nucleares (cuantificadores, adverbios que modifican a un adjetivo) y no incluye coordinación, tratando sólo segmentos nominales (cobertura: 98,5 %; precisión: 95%). (*K. Church* 1988) tampoco contempla la coordinación de estructuras simples (segmentos nominales precisión 98,6%). Por el contrario, (*L. Ramshaw & M. Marcus* 1995) emplea técnicas de transformación a la manera de *Brill* y que sí que incluye los segmentos nominales compuestos por coordinación y segmentos comparativos, pero excluye el genitivo sajón; por lo tanto, no son comparables sus cifras (cobertura: 93,5 %; precisión: 93,1%), obtenidas además sobre un corpus segmentado automáticamente.

Los parámetros empleados suelen ser contextos de etiquetas, aunque (*L. Ramshaw & M. Marcus* 1995) emplea también patrones de palabras, quizá por un inadecuado etiquetado morfosintáctico.

2.2.2.3 Corpus y bases de datos sintácticos en castellano

Parte del corpus *LEXESP* ha sido etiquetado sintácticamente (75.000 palabras) con 7 etiquetas (sintagma nominal, sintagma Verbal, sintagma preposicional, sintagma adjetivo, sintagma adverbial, infinitivo y nexo subordinante). De manera similar ha sido analizado el corpus *CPirápides*, que consta de 4.970 oraciones sencillas (31.376 palabras en total, de longitud máxima inferior a 15 palabras) y con un número bajo de elementos adjuntos (*M. Civit & I. Castellón* 1998).

La *Base de datos sintácticos del español actual* contiene gran cantidad de información sobre rección y estructura verbales en castellano que pueden resultar muy útiles para analizar de modo no parcial (*G. Rojo et al* 2001).

2.2.2.4 Sistemas de análisis sintáctico en castellano

El sintetizador *Boris* dispone de un conjunto de reglas heurísticas (sin diccionario exhaustivo) destinadas a detectar posibles puntos para la inserción de pausas, puntos que se corresponden con posibles límites externos o internos de sintagmas (*J.M. Pardo et al* 1995).

En (*H. Jiménez & G. Morales* 2001) se emplean árboles de decisión y ganancia de información normalizada para detectar sintagmas nominales con una cobertura superior en torno al 98 % con un 97 % de precisión.

El sistema APOLN (*A. Molina et al* 1999) emplea una secuencia de autómatas finitos con concordancias y la estrategia *longest-match;* en una evaluación sobre *CPirápides* alcanzó tasas de cobertura superiores al 99% (salvo en sintagmas adverbiales: 95,2 %) y tasas de precisión superiores al 94 % (salvo en sintagmas adjetivales 66,7 %).

Al evaluar sobre parte de LEXESP (corpus sintácticamente más complejo), las cifras bajan (en el peor de los casos, la precisión de los sintagmas adjetivales baja al 77,6% aunque aumente la cobertura hasta un 80,9 %).

En la presente Tesis se empleará el *parser CYK* descrito en (*J.M. Montero* 1992). Aunque este analizador requiere que las reglas estén escritas en Forma Normal de *Chomsky (FNC)*, existe un algoritmo que permite convertir una gramática de contexto libre a FNC. En (*G. Erbach* 1994) se generaliza el algoritmo para gramáticas en formato libre diferentes al FNC.

Se ha optado por el empleo de reglas de contexto libre, dado que son equivalentes a las reglas basadas en rasgos no recursivos, aunque tengan menos elegancia descriptiva[8].

2.3 Análisis y modelado prosódico

La prosodia está constituida por los aspectos rítmicos y entonativos del habla, y se encuentra relacionada con fenómenos físicos segmentales tales como la duración, la frecuencia fundamental y la intensidad, y con fenómenos lingüísticos tales como el acento o la división en grupos fónicos. Afecta, por tanto, a dominios superiores al segmento (fonema, sílaba...), apoyándose sobre magnitudes segmentales e intra-segmentales que varían a lo largo del tiempo. La importancia de la prosodia radica en que es uno de los principales mecanismos de que dispone el ser humano para transmitir intención, expresar actitudes o emociones, crear el foco de un discurso oral, llamar la atención, etc. (aunque no es el único, como veremos a lo largo de esta Tesis).

Así como fonéticamente existe un conjunto de unidades discreto y amplio, pero bien estudiado, que permiten transcribir una elocución, no existe un conjunto equivalente de unidades prosódicas.

2.3.1 Entonación y F0

La **frecuencia fundamental** (F0) de un fragmento de habla es el primer armónico que presenta su espectro y que se corresponde, desde el punto de vista de producción de habla, con la periodicidad de la vibración de las cuerdas vocales. Estas magnitudes están relacionadas con el *pitch*, que es la percepción del tono en sonidos complejos que, aunque viene influido principalmente por dicha frecuencia fundamental, también experimenta el influjo de los armónicos superiores por debajo del primer formante, siendo posible que ruidos de espectro plano pero estrecho sean percibidos como dotados de *pitch*[9].

Desde un punto de vista lingüístico, la **entonación** es la línea melódica con que se pronuncia un mensaje (*J. Alcina & J.M. Blecua* 1975). Entre las funciones que pueden desempeñar la entonación y los distintos patrones entonativos destacan los siguientes:

- **Integrar o delimitar frases o sintagmas**: así las expresiones explicativas o parentéticas suelen ser marcadas por medio de pausas y una entonación especial.
- **Desambiguar**: el acento léxico contribuye a distinguir pares mínimos como las palabras /izo/ e /*izó*/.
- **Aportar información sobre la modalidad oracional**: la entonación de las oraciones interrogativas y de las enunciativas es diferente, informando al oyente sobre qué modalidad es la que desea transmitir el locutor.
- **Resaltar elementos del discurso**: por medio de picos o valles entonativos es posible destacar una o varias palabras dentro de una elocución.
- **Transmitir un estado de ánimo, una emoción o una actitud**: la entonación se ve notablemente influida por el estado de ánimo que tiene o simula tener el locutor, siendo posible identificar el estado de ánimo por medio de la entonación, en algunos casos.
- **Revelar el origen sociolingüístico**: la entonación habitual de una persona también puede revelar su origen geográfico o social, incluso su edad.

2.3.1.1 Escuelas de análisis de contornos de F0

Internacionalmente, en el análisis de contornos de frecuencia fundamental existen 2 escuelas principales (*A. Botinis et al* 2001):

§ **La fonológica**: la frecuencia fundamental se describe por medio de una serie de unidades abstractas, a partir de las cuales se presentan una serie de niveles discretos y diversas transiciones.

§ **La acústico-fonética**: describen la curva de entonación como un continuo con complejos patrones de inflexiones o movimientos de F0 (IPO)

2.3.1.1.1 Modelo TOBI (Tone and Break Indices)

Se trata de un modelo fonológico originado por los trabajos de *J. Pierrehumbert*, *M. Beckman* y *J. Hirschberg* fundamentalmente (*M.E. Beckman & G.M. Ayers* 1994). Analiza cada contorno de F0 de un modo jerárquico como una serie de eventos de *pitch* (no como una secuencia de puntos clave y transiciones interpolables entre los mismos) y una serie de índices de separación entre palabras. Los índices pueden ser:

§ **Nivel 0**: separación entre un verbo y sus pronombres clíticos.

§ **Nivel 1**: el que separa la mayoría de las palabras.

§ **Nivel 2**: se produce una pausa entre palabras, pero no afecta a la curva de tono.

§ **Nivel 3**: no se produce una pausa, pero la curva de F0 se resetea a nivel alto o a nivel bajo, dependiendo de la declinación producida (*downstepping*).

§ **Nivel 4**: pausa con *reset* de F0.

Por su parte los eventos de tonos pueden estar asociados a:

§ **Un acento**: eventos de nivel alto para el locutor en cuestión (H*), de nivel bajo o medio (L*), de subida abrupta (L+H*), nivel alto con *downstep* (¡H)...

§ **Un límite de grupo entonativo (phrase, entre dos reset)**: *reset* a nivel bajo (L-) o a nivel alto (H-), un *plateau* final (H- L%).

§ **De grupo fónico (boundary, entre 2 pausas)**: final en nivel bajo (L%), en nivel alto (H%), subida de continuación (L- H%), inicial en nivel alto (%H), subida final en interrogativas (H- H%), etc..

Se intenta describir subjetiva y perceptualmente el contorno de F0 pero sin interpretarlo, y tiene gran interés lingüístico. El etiquetado manual ToBI resulta muy costoso en tiempo de personal experto y resulta difícil su automatización (*M. Ostendorf & K. Ross* 1997) o su semiautomatización (*A. Syrdal et al* 2001). En castellano existe una descripción cualitativa de este tipo en (*J. M. Sosa* 1999).

2.3.1.1.2 Modelos acústicos

El **modelo del IPO** es un modelo acústico no jerárquico desarrollado en el *Institute for Perception Research* y se basa en la estilización[10] y estandarización de contornos, en un conjunto de movimientos básicos (definidos por sus rasgos: anchura, velocidad, rango, dirección y posición) y en 3 líneas de declinación (alta, media y baja).

En el modelo IPO se inspira el **modelo de Garrido**, que es un modelo jerárquico que también tiene estilización de contornos, 3 líneas de declinación (alta media y baja), 3 niveles de F0 (alto, medio y bajo, donde el nivel medio tiene en cuenta los picos de F0 excepcionalmente bajos, o los valles altos), 3 movimientos de F0 (ascendente, descendente y plano) y la estilización por tramos rectos. Estructura la entonación en varios niveles, entre los que destacan el de párrafo, el de oración y el de grupo entonativo, definiendo la existencia de declinación (supra-líneas de declinación) en el nivel de oración (no sólo en el grupo entonativo), así como 3 partes dentro de todo grupo entonativo en castellano: inicial, intermedia y final, además de un contorno especial para los casos de una sola tónica (*J.M. Garrido* 1996).

El **modelo TILT** analiza la curva de F0 como una serie de acentos y tonos límite (*boundary tones*) caracterizados por un conjunto de parámetros de amplitud, posición, duración y forma, fácilmente computable a partir de la curva de F0 (*P. Taylor* 2000).

El **modelo INTSINT** también se basa en una serie de marcas de subida, bajada, máximo y mínimo, interpolados por medio de *splines*, susceptible también de análisis automático (*J. Veronis et al* 1998)

2.3.1.1.3 Modelos aditivos

El **modelo de Fujisaki** emplea un modelo jerárquico aditivo de la curva de F0, que es cuantitativo, paramétrico y continuo en el tiempo, y que ha sido empleado para modelar la entonación de diversos idiomas incluido el español (*H. Fujisaki, S. Ohno et al* 1994), con resultados comparables a los de las mejores técnicas de modelado (*J. Gutiérrez-Arriola* 2001b).

Según el modelo de Fujisaki (*H. Mixdorf* 1998), el contorno de F0 se compondría de 3 componentes aditivas en el dominio logarítmico de F0:

- **Un nivel base**: dependiente del locutor o dependiente de la frase.
- **Una o varias componentes de grupo entonativo (phrase)**: pulso muy asimétrico capaz de modelar fenómenos como la declinación o anti-declinación, el foco ancho, etc.
- **Una o varias componentes acentuales**: pulso asimétrico capaz de modelar la acentuación, o la anti-cadencia de continuación.

La complejidad de la formulación del modelo de Fujisaki dificulta enormemente el análisis y la extracción automática a partir de mediciones de F0, que resulta muy costosa en tiempo y con numerosas revisiones manuales. El método más habitual es el de análisis por medio de síntesis (*analysis-by-synthesis*) en el que se realiza una búsqueda completa usando pasos cuantificados, dentro del dominio razonable de cada parámetro, hasta alcanzar la combinación de valores que mejor se ajusta al contorno medido[11].

Como es un modelo cuantitativo, puede ser empleado para modelar la curva de F0 en un sistema de síntesis de voz, y cada vez abundan más los métodos entrenados de predicción de comandos: técnicas conexionistas (*K. Hirose, M. Eto, N. Minematsu & A. Sakurai* 2001), árboles de regresión (*H. Mixdorf* 1998) o el vecino-más-próximo o *Nearest Neighbour* (*J. Gutiérrez-Arriola* 2001b).

El **modelo de** *Bell Labs*, también aditivo, sustituye la formulación matemática por patrones escalables en el tiempo (*templates*) que modelan las curvas de acento, de frase y de segmento (*J. Van Santen & B. Moebius* 1997).

2.3.1.2 Acentuación y desacentuación léxica. Foco

El **acento léxico** (*stress*) es la mayor prominencia prosódica de un segmento (fonema, sílaba) dentro de una unidad mayor (palabra, oración). La prominencia percibida puede deberse a mayor duración, F0 o intensidad o a una combinación de ellas (*G. Fant et al* 2001), aunque también tiene una componente de fonación (mayor *OQ*, menor *tilt*, mayor *skew*).

Todos los modelos antes mencionados tratan el fenómeno de la acentuación, pues presenta una alta correlación con los picos del contorno de F0, aunque las **palabras función** (determinantes, clíticos, etc.) suelen ser sistemáticamente **desacentuadas**, y también pueden desacentuarse algunas **expresiones numéricas u horarias**, **verbos auxiliares**, etc. De la misma manera, el acento en los nombres compuestos recae habitualmente en una sola de las palabras, la que el locutor considera principal. También la mayor o menor longitud del grupo fónico influye sobre la mayor o menor presencia de acentos léxicos. Aunque podría parecer que en estructuras nominales (número + sustantivo o sustantivo + adjetivo), es el núcleo nominal el que lleva la prominencia siempre, y es el complemento quien se ve desacentuado a veces (*H. Mixdorf* 1998), este hecho se puede invertir debido a factores semánticos (acento contrastivo: resaltar una determinada información por oposición a otra; acento de novedad: resaltar una información nueva) o personales espontáneas (eufonía o familiaridad).

No conviene confundir el acento léxico con el focal, aunque muchos de los acentos focales coincidan con acentos léxicos. Lingüísticamente el **foco** es el elemento (foco estrecho) o los elementos (foco ancho) que semánticamente son más relevantes dentro de una elocución, y que reciben por ello una prominencia especial superior a la del acento (se puede hablar también de que son acentos enfáticos). Suele contener información nueva o de contraste (respecto al contexto actual o respecto a la elocución anterior), aunque por su naturaleza semántica resulta difícil su procesamiento automático. La presencia de un foco cercano puede provocar la desacentuación de los elementos

previos o siguientes (si se trata de un foco estrecho) o el realce de los mismos si el foco es más ancho y de transición más lenta (*Y. Xu* 1999)

Pero no sólo la desacentuación de un acento léxico y la presencia de una acento focal pueden afectar a la correlación entre presencia de un acento léxico y la aparición de un pico de F0, también puede hacerlo un **desplazamiento del pico de f0**. En (*J. Llisterri et al* 1995) se expone como habitual (70 %) que en las oraciones enunciativas se produzca un retraso del pico de F0 de la sílaba acentuada a la siguiente sílaba. Este desplazamiento se ve condicionado por barreras sintácticas (no cruzar de un sujeto a un verbo, no acercarse o cruzar una pausa) o por factores rítmicos individuales (como es lógico, desplazar el pico de F0 modifica la melodía de la frase, lo cual puede tener consecuencias eufónicas).

La predicción del acento en textos sin restricciones es un tema de investigación de gran importancia para el modelado prosódico (*R. Sproat et al* 1992). Parámetros que influyen en la acentuación son:

- **Distancia de la palabra al comienzo y fin de oración**.
- **Distancia a las pausas anterior y siguiente**.
- **Número total de palabras en la oración**.
- **Clases amplias a la que pertenece la palabra**: clase cerrada acentuada, clase cerrada desacentuada, clase cerrada clítica y clase abierta.
- **Elementos antepuestos** : adverbios, sintagmas preposicionales, etc.

Para casos especiales (como nombres compuestos propios o no) suelen emplearse reglas *ad hoc (A. Syrdal et al* 2001).

2.3.1.3 Relaciones entre F0, intensidad y duración

En posiciones pre-pausa abundan las alteraciones de los valores habituales de los parámetros relacionados con la prosodia, esto es: F0, duración e intensidad. Si las oraciones enunciativas se suelen caracterizar por contornos de F0 de final descendente, este descenso suele venir acompañado de un incremento de la duración y un descenso de la intensidad que parecen revelar una estrategia general de reducción del esfuerzo vocal y articulatorio (*R. Herman* 1996).

2.3.1.4 Micro-prosodia o micro-melodía

Algunos autores señalan que cada vocal tiene un valor de F0 intrínseco que hace aumentar o disminuir el valor local de F0 (las vocales *high* tendrían mayor F0 intrínseca) y que el contexto segmental (fonemas oclusivos sordos o nasales) puede también hacer variar localmente la curva de F0. Sin embargo estos fenómenos no son apreciables en curvas suavizadas por el modelo de Fujisaki (*H. Mixdorf* 1998), por curvas de Bézier (*D. Escudero-Mancebo et al* 2002) o basadas exclusivamente en información de la F0 en sílabas (*J.A. Vallejo* 1998).

2.3.1.5 Relaciones entre entonación y sintaxis

Es conocido que las relaciones entre la estructura sintáctica y la estructura entonativa no son simples sino complejas y muy variables. Ciertas estructuras están relacionadas con la posibilidad de contornos entonativos determinados (elementos explicativo o parentéticos, enumeraciones, elementos adelantados a posición temática, elementos subordinados, estructuras origen-destino, etc.) e influyen, cuando no condicionan, la entonación.

Asimismo la introducción de una pausa allí donde se rompen ligaduras sintácticas de una cierta fuerza, provoca que el locutor genere una entonación ascendente de continuidad en grupos fónicos enunciativos.

2.3.1.6 Patrones entonativos en castellano

En castellano destacan los estudios lingüísticos (*N. Tomás* 1948), (*A. Quilis* 1981) y (*J.M. Garrido* 1991) así como, desde un punto de vista de Tecnologías del Habla, (*J.C. Olabe* 1983) y (*J.M. Pardo et al* 1987*).*

A la hora de caracterizar los distintos patrones es casi unánime la división del contorno en 2 **zonas terminales** (inicial y final) y 1 **zona no terminal** (intermedia), siendo la zona terminal final la que se considera más importante, sobre todo a la hora de delimitar e integrar sintagmas o para definir la modalidad oracional (*J.M.*

Garrido 1991). Estas zonas vienen definidas por la posición de la primera y la última tónica, como parecen confirmar (*J.M. Garrido* 1991) y (*J.A. Vallejo* 1998). En la definición de la modalidad oracional es el tramo final el más relevante, como sucede en otros idiomas (*H. Mixdorf* 1998).

Otro elemento habitual en las descripciones de patrones es la **declinación,** progresivo descenso de los picos de F0 a lo largo de un grupo fónico. Estudios previos para el español mejicano como (*P. Prieto et al* 1996) revelan que más de un 65 por ciento de los valores de los picos de F0 se pueden predecir en función del pico anterior.

Finalmente es posible encontrarse con *resetting* de la curva de F0 que no coinciden con pausas del locutor, generando varios grupos entonativos dentro de un mismo grupo fónico.

Los parámetros físicos que ayudan a describir cuantitativamente los patrones (*J.M. Montero et al* 2002) son:

- **Rango**: es habitual que se asocie un mayor rango con emociones tales como la alegría.
- **Tono medio o registro**: una bajada general de la curva de F0 suele asociarse con situaciones anímicas como la tristeza.
- **Pendiente de un movimiento de F0**: ascendente o descendente. Los movimientos de gran pendiente suelen provocar un realce o prominencia del elemento al que está asociado, colocando el elemento en una posición de foco.
- **Valor máximo**: es conocido que los máximos valores de F0 se dan en situaciones emocionales extremas como las de sorpresa o pánico.

Siguiendo la exposición de (*J.M. Garrido* 1991) podemos clasificar los patrones del castellano analizando las distintas modalidades oracionales en castellano:

- **Enunciativos**: suelen presentar declinación y un contorno final o tonema descendente (también denominado cadencia), aunque si se produce una pausa forzada se puede producir una subida (o anti-cadencia) que señala que no se trata de un final de oración, sino que habrá una continuación.
- **Interrogativos (absolutos, relativos o pronominales)**: preguntas cuya respuesta es del tipo *si o no*. Es típico su final ascendente o anticadencia y, en el caso de las pronominales, un pico inicial mayor.
- **Exclamativos (pronominales o no)**: suelen presentar mayor rango y mayor número de picos. Muchas veces no presentan una estructura oracional completa.
- **Volitivos**: imperativo, desiderativo, etc.

A lo largo de esta Tesis veremos ejemplos de enunciativas e interrogativas absolutas (dentro de un dominio restringido) y exclamativas no pronominales (base de datos de voz con emociones). Hay que señalar que la modalidad oracional también puede venir marcada, además de por el patrón entonativo, por la presencia de determinadas partículas (interrogativos y exclamativos), por el orden de las palabras, etc.

2.3.1.7 Definición y diseño de una base de datos prosódica

Para el estudio de los fenómenos prosódicos en conversión texto a voz se suele comenzar por la definición y grabación de una base de datos, cuyo texto es leído o interpretado por un locutor, frecuentemente profesional. La fase de definición tiene por objetivo contemplar un importante número de variantes segmentales y prosódicas que sean representativas y permitan un análisis dentro del dominio planteado.

Una vez grabada la base de datos es necesario marcar los elementos que serán objeto de estudio: duración, F0, marcas glotales, formantes… Hoy en día proliferan cada vez más los métodos automáticos y semiautomáticos de marcado (*M.J. Makashay et al* 2000) (*D. Torre* 2001).

La grabación de bases de datos espontáneas podría ser de gran importancia para dotar de mayor variedad a los sintetizadores, pero los primeros prototipos de voces comerciales se realizan con bases de datos de contenidos más controlados y mejor orientados a la aplicación deseada.

Dependiendo del objetivo, las bases de datos prosódicas pueden ser de uno o varios locutores. En general, para llevar a cabo un modelado de una voz sintética, es preferible emplear un solo locutor, ya que el objetivo es que el

sistema lo imite (*J.A. Vallejo* 1998), mientras que si el objetivo es un análisis general del fenómeno entonativo, es preferible emplear varios locutores (*J.M. Garrido* 1991).

Respecto a los textos empleados, la división no es tan clara. Si bien para analizar es preferible emplear frases de laboratorio o textos muy controlados, que permiten definir pares mínimos diferenciados por un único parámetro (*J. M. Garrido* 1996) o incluir fenómenos poco frecuentes; también al modelar es necesario añadir frases de laboratorio a las frases que forman un discurso *real* de naturaleza oral (*J.A. Vallejo* 1998).

La dicotomía entre habla leída o espontánea o entre un locutor profesional o no profesional suele decantarse por el habla leída y el locutor o locutora profesional, especialmente si se intenta modelar un servicio interactivo como lo haría un profesional (*J.M Montero et al* 2000), o si se intenta modelar la simulación de situaciones emocionales (*J. M. Montero et al* 1999).

Entre los parámetros lingüísticos que influyen en el texto del corpus podemos destacar: número de sílabas de cada grupo fónico, distancia entre acentos, límite final del grupo fónico (*J.M. Garrido* 1996), etc.

La selección de textos para una base de datos suele ser un problema de búsqueda en un espacio de dimensiones combinatorias. La solución óptima (*full search*) es computacionalmente intratable, aunque pueden encontrarse soluciones sub-óptimas que automaticen el proceso mediante técnicas como los algoritmos voraces (*P. van Santen et al* 1997a), o los algoritmos genéticos (*O. Boëffard et al* 1997).

En (*P. van Santen et al* 1997a) se describe cómo conseguir un conjunto mínimo de textos que contengan todos los fonemas al menos una vez. Para ello los autores usan pesos inversamente proporcionales a la frecuencia de cada fonema de la frase, a fin de conseguir seleccionar antes aquellos fonemas menos frecuentes; una vez incorporado un fonema al conjunto seleccionado, su peso pasa a valer 0.

En (*J. Shen et al* 1999) se aplica la selección a la definición de una base de datos para reconocimiento de voz. Se intenta escoger un conjunto mínimo de frases cuya distribución fonética sea lo más parecida que se pueda a un conjunto de frases mucho mayor. Su algoritmo se compone de 2 fases:

- *Covering phase***:** tiene por objeto conseguir un conjunto mínimo que contenga todos los fonemas, al menos una vez, de manera similar a lo descrito en (*P. van Santen et al* 1997a).
- *Distribution phase:* busca conseguir la distribución de probabilidad deseada; la distancia de selección empleada es el producto escalar de los vectores de cada frase y el vector objetivo final, normalizado por el producto de los módulos de los mismos (en cada paso se elige la frase que fonéticamente más se parece a la distribución no cubierta todavía). Tras cada paso de la selección se actualiza el vector objetivo.

Aunque la técnica es interesante, sólo se expone una medida directa sobre la calidad de los distribución obtenida: una correlación normalizada de 0,8742, frente a los 0,6602 obtenidos eligiendo algunos párrafos.

En esta Tesis se expondrá un método voraz (*greedy)* de diseño de bases de datos relacionado con los anteriores, pero orientado a conseguir resumir una base de datos en un número máximo de ejemplos posibles, empleando una aproximación escalonada al objetivo final.

2.3.1.8 Métodos para la generación de curvas de F0

Pueden ser paramétricos o no paramétricos, generales o específicos.

2.3.1.8.1 Modelos paramétricos de patrones entonativos

Se basan en la sucesión de tramos rectos ascendentes y descendentes que se asignan en función de la posición de las vocales tónicas y átonas y del tipo de frase. Los valores de estos picos y valles se obtienen de un análisis estadístico de la base de datos. Ejemplos se pueden encontrar en (*P. Moreno et al* 1989) (*E. López-Gonzalo* 1993), (*M.A. Rodríguez et al* 1993), el modelo *Hat* (*D. Klatt* 1987*): 360-363*), el modelo *Rise / Fall / Connection* (*P. Taylor* 1994) y el modelo *TILT* (*P. Taylor* 2000). Pueden resultar imprescindibles en casos de escasez de datos y será una de las alternativas de modelado que emplearemos en la presente Tesis.

2.3.1.8.2 Métodos no paramétricos basados en memoria

Aunque se basen en la acumulación de ejemplos concretos de parámetros de entrada y su correspondiente solución, poseen cierta capacidad de generalización. En el momento de sintetizar la prosodia se limitan a buscar el ejemplo más parecido que recoja la base de datos y aplicarlo (*W. Daelemans* et al 1999).

2.3.1.8.3 Modelos conexionistas

Las redes neuronales artificiales están formadas por un número elevado de unidades sencillas de procesamiento en paralelo, lineales o no lineales, altamente interconectadas por medio de pesos. Aunque algunos autores han optado por redes neuronales recursivas (*C. Traber* 1992), los perceptrones multicapa se han revelado como muy efectivos a la hora de predecir la curva de tono (*J.A. Vallejo* 1998*)*, si se modela la continuidad por medio de enventanado de algunos de sus parámetros de entrada (*S. Tournemire* 1997) (*Y. Morlec et al* 1997).

Los buenos resultados de (*J.A. Vallejo* 1998) hacen que la técnica conexionista sea la que empleemos en el modelado de F0 en un dominio restringido.

2.3.1.8.4 Métodos simbólicos inducidos automáticamente

Se basan en algoritmos genéticos (*J.A. Vallejo* 1998). Los parámetros relevantes se determinan por medio de algunas de las técnicas anteriores, y buscan fórmulas que modelen la curva de F0 por medio de la combinación de dichos parámetros y un conjunto finito y definido de operadores.

La ventaja de los métodos simbólicos inducidos está en que proporcionan información simbólica explícita comprensible por expertos humanos, con la posibilidad de ser entrenados e inferidos automáticamente.

2.3.1.9 Percepción de la frecuencia fundamental

La capacidad de discriminación de frecuencias del oído establece qué diferencias es capaz de percibir cuando se le presentan distintos sonidos consecutivamente (*jnd*: *just noticeable differences*). Realizando medidas con sinusoides puras, esta capacidad está en torno a 1 Hz para frecuencias inferiores a 500 Hz (*E. Zwickerl* 1990). Sin embargo, la percepción del tono en sonidos más complejos es un fenómeno asimismo más complejo. Aunque en (*D. Klatt* 1973) se informa de discriminaciones de 0,3 Hz con habla sintética de F0 constante, en (*J. Pierrehumbert* 1979), donde se trabaja con discriminación de contornos de F0, se informa de que la *jnd* estaría entre 7 y 12 Hz en torno a 120 Hz de F0 media; por su parte, en (*M.S. Harris et al* 1987) se sitúa el *jnd* de la voz natural entre 5 y 16 Hz, mientras que en (*J. t'Hart* 1974) se habla de experimentos de percepción de movimientos ascendentes o descendentes de F0 entre 1,5 y 4 semitonos. En (*J.M. Garrido* 1991) se menciona (sin aportar datos empíricos de evaluación) que, aunque en general 10 Hz es un buen umbral para detectar inflexiones de F0 perceptualmente relevantes, existen casos que impiden tomar ese umbral como absoluto.

En todo caso, parece lógico que la robustez que caracteriza la comunicación oral humana debe hacer que las diferencias que transmiten información (las interesantes para la síntesis de voz) deben encontrarse bastante por encima del umbral de percepción o discriminación.

La representación auditiva de las frecuencias de los sonidos no es lineal; además del umbral antes definido, el oído responde de manera aproximadamente logarítmica, especialmente para frecuencias altas. Así la escala musical en octavas responde a la fórmula log2(f/127.09), mientras que la escala **mel** es 1127*ln(1+f/700). En (*H. Traunmüller et al* 1995) se informa de experimentos donde oyentes juzgaban como equivalentes aquellos intervalos de F0 que eran iguales en semitonos.

En experimentos con voz femenina y masculina se ha observado que se necesitan mayores excursiones de F0 para que la voz femenina produzca impresiones de prominencia similares (*C. Gussenhoven et al* 1998), y que los picos de F0 son más importantes que los valles en la entonación de idiomas no tonales (*T. Portele et al* 1997).

2.3.1.10 Normalización de valores de F0

Aquellas escuelas que describen los contornos de F0 como una secuencia de niveles discretos destacan la naturaleza relativa de estos niveles (*RAE* 1973). En (*J. Pierrehumbert* 1987) se refieren los valores de F0 a una hipotética línea básica dependiente del hablante, de tal manera que el fundamental normalizado es el resultado de restar el valor de la recta básica y dividir por ese mismo valor. Otros autores citados en (*J.M. Garrido* 1991) normalizan por el valor mínimo y el rango de un hablante, aunque para evitar normalizar por valores

excepcionales, podría ser más interesante aplicar un *z-score*, esto es, normalizar por la media y el rango que incluye el 95 por ciento de los valores, de manera similar a lo que se realiza en el modelado de duraciones (*R. Córdoba et al* 1999).

2.3.1.11 Evaluación del modelado de F0

Para la **evaluación** del modelado de F0 se han propuesto diversos métodos, tanto objetivos como subjetivos. Dentro de los **métodos objetivos** tienen especial relevancia los más simples, basados en la medida del error cuadrático medio o en el error absoluto (esto es, basados en la distancia entre la F0 generada automáticamente y la F0 medida en la base de datos, para cada uno de los fonemas sonoros donde tiene sentido definir y calcular la F0). Presentan la ventaja de su facilidad de cálculo, pero no tienen en cuenta efectos de percepción que pueden hacer que sus resultados sean inexactos en ciertos casos. Dentro de los **métodos subjetivos** encontramos los basados en evaluación de expertos (que evalúan la calidad de la curva en sí) y los basados en evaluación por parte de oyentes no expertos (basados en la aceptación dentro de una escala de 5 puntos como la del MOS –Mean Opinion Score- o basados en la comparación y selección entre un par de alternativas).

Es habitual que los oyentes prefieran las voces de mujer para las aplicaciones IVR, aunque la inteligibilidad puede ser algo menor (*L. Aguilar et al* 1994).

Prueba de inteligibilidad de palabras o campos clave: aunque normalmente este tipo de pruebas se realizan con frases sintácticamente correctas pero semánticamente anómalas (*L. Aguilar et al* 1994), en un caso de dominio restringido sería más adecuado un conjunto de palabras de características fonéticas equilibradas que los oyentes deben identificar y transcribir.

2.3.2 Duración y ritmo

El **ritmo** es un fenómeno supra-segmental que se apoya sobre las duraciones de los segmentos que define una curva prosódica cuyo valor medio denominamos **velocidad de elocución** (número de fonemas por segundo que contiene una elocución). Se trata de una magnitud de difícil medición, dado que la duración es un fenómeno muy variable, donde los distintos segmentos poseen duraciones intrínsecas condicionadas por el contexto.

Las desviaciones del ritmo se pueden medir por medio de *Pairwise Variability Index* (*D. Gibbon et al* 2001), que es el promedio de las diferencias entre duraciones de vocales o sílabas sucesivas, en valor absoluto, normalizadas por la media de las duraciones de las unidades.

Si se dispone de oraciones de referencia (por ejemplo, si se dispone de una frase emocionalmente neutra que comparar con oraciones tristes, alegres, etc.), es posible alinearlas mediante DTW, obteniéndose el ritmo como derivada temporal del alineamiento DTW (*S. Ohmo et al* 2001).

2.3.2.1 Normalización de la duración

Muchos son los factores que condicionan la duración de un segmento de voz, y cuyo efecto puede ser compensado aplicando normalizaciones a la duración (*D. van Kujik & L. Boves* 1999):

- Normalizar por la duración media de cada segmento de la frase.
- Normalizar por la duración media y la varianza de cada segmento de la frase.
- Normalizar por la velocidad media de vocales y consonantes de la frase.
- Normalizar por la velocidad media de los fonemas de la frase.
- Normalizar por la duración de las vocales precedentes.
- Normalizar por la duración de las vocales dentro de una ventana temporal

Algunos de estos métodos de normalización buscan modelar las duraciones de cada unidad independientemente de la velocidad de elocución, lo cual supone no modelar el ritmo global, mejorando las tasas de acierto de las duraciones.

2.3.2.2 Modelos de duraciones

2.3.2.2.1 Modelo Multiplicativo-aditivo

Tiene su origen en el modelado de Klatt (*D. Klatt* 1987), que establecía una duración mínima para los fonemas, una duración intrínseca para cada fonema y un factor de modificación de la misma dependiente del contexto fonético (alargamiento), la posición en la frase (alargamiento pre-pausa), el acento, la velocidad de elocución, el tamaño de la palabra (mayor longitud implica menores duraciones de los segmentos), el tipo de sílaba (grupos consonánticos, abiertas o cerradas), etc. El modelo multiplicativo se podría llamar aditivo si trabajamos en un dominio logarítmico. Sin embargo, esto sigue sin permitir el modelado de las relaciones entre parámetros que no son independientes; por ejemplo, las vocales no acentuadas pueden tener duraciones similares a las acentuadas en palabras aisladas si hay bastante alargamiento pre-pausa (*J. Sánchez* 2000), debido a que el alargamiento pre-pausa parece afectar más a las vocales no acentuadas (*J. Van Santen* 1992). Todo ello se puede paliar mezclando multiplicaciones y adiciones, complicando con ello su estimación.

2.3.2.2.2 Modelo conexionista o neuronal

La dificultad para encontrar un modelo de las duraciones ha hecho muy atractivo el empleo de las redes neuronales (*N. Campbell* 1992). Como siempre en estos casos, se trata de investigar qué parámetros producen mejoras en el modelado (*R. Córdoba* 1999).

2.3.3 Pausado

Es habitual que una oración, especialmente si es algo larga, se divida en varios grupos fónicos separados por pausas. Estas pausas pueden tener un origen gramatical, y muchas veces están señaladas en las versiones textuales (si existen) por medio de signos de puntuación (**pausas gramaticales**):

- **Subordinadas explicativas, expresiones parentéticas, subordinadas de relativo anidadas o largas**.
- **Subordinadas adelantadas a la principal u oraciones principales muy largas**.
- **Coordinación entre elementos de una cierta longitud.**
- **Elementos extraídos de su posición habitual o dotados de gran libertad**: vocativos, interjecciones, saludos, adverbios, preguntas de confirmación o rechazo, sintagmas preposicionales, etc.
- **Ambigüedades estructurales**: debidas a que existan varias posibilidades de rección, típicamente entre sintagmas preposicionales que complementan a sintagmas nominales o verbales.
- **Focalización**: de un elemento nuevo o contrastivo.

Si la oración es larga, es frecuente que el locutor precise de la realización de **pausas respiratorias espontáneas**, pausas no impuestas por motivos sintáctico-semánticos, pero sí condicionadas por la estructura sintáctica y semántica de la oración.

Finalmente, los discursos espontáneos contienen dudas y rectificaciones (*speech repairs*), que provocan **pausas espontáneas de duda o rectificación**. En (*A. Batliner et al* 1998) se detectan y etiquetan manualmente hasta 59 tipos de fronteras sintácticas que son posibles (o imposibles) posiciones de pausa en diálogos.

Variables que influyen sobre las pausas espontáneas o respiratorias son (*M. Wang & J. Hirschberg* 1992):

- **Distancia al comienzo o fin de frase:** posiciones muy poco probables para una pausa.
- **Contexto de palabras o etiquetas morfosintácticas**: es poco frecuente romper un sintagma nominal.
- **Acento de la palabra.**
- **Posible homologación entre palabras.**
- **Distancia al comienzo o fin de frase** normalizada por la longitud del anterior grupo fónico.
- **Longitud de la frase**: en tiempo o en sílabas y en palabras.

- **Tipo de sintagma y puntuación**.
- **Modalidad de oración**.
- **Pertenencia a un sintagma nominal, distancia desde su comienzo y tamaño del sintagma.**

En castellano destacan los estudios de (*J.M. Pardo et al* 1987), basado en reglas, de (*D. Casacuberta et al* 1997) y de (*J. Hirschberg & P. Prieto* 1996).

Las bases de datos suelen ser textuales y etiquetadas por un experto lingüista, más que basadas en una base de datos de voz.

En el modelado predominan los sistemas por reglas (*M. Rodríguez et al* 1993) y los basados en *CART* (*P. Taylor* 1998). La evaluación puede ser totalmente objetiva (número de errores sobre una base de datos) o subjetiva (un experto clasifica los errores como aceptables o no).

2.4 Personalización de voz y habla con emociones

Aunque las emociones son casi tan antiguas como la vida animal y han sido estudiadas por biólogos y filósofos desde tiempos de Aristóteles hasta Sartre pasando por Darwin, el interés por las emociones humanas como parte importante de los procesos de decisión, inteligencia y memoria ha crecido considerablemente en los últimos años, alcanzando incluso el ámbito de lo popular (*D. Goleman* 1995).

Aunque por su propia naturaleza las emociones son difíciles de capturar en una definición, algunos rasgos se encuentran presentes en las de varios autores (*D. Casacuberta* 2000):

- Se trata de **estados mentales** de los animales y los seres humanos, conscientes o no, de una cierta intensidad y duración breve.
- Pueden actuar como catalizador, inhibidor, favorecedor u obstaculizador de las reacciones humanas ante determinados eventos externos o internos (**función de planificación y adaptación**), eventos que pueden ser clasificados como beneficiosos o dañinos (**función valorativa o** *appraisal*).
- Pueden provocar ciertas alteraciones fisiológicas prototípicas (*bodily changes*), perceptibles desde el exterior (**función comunicativa o social-cultural**)[12].

Estas características permiten diferenciar las emociones de los estados de ánimo (más duraderos, menos intensos) o las actitudes (no reactivas), pero lo que más nos interesa en esta Tesis es que las emociones pueden conllevar patrones de cambio en la voz y que suelen tener una función comunicativa (que informa a los demás del estado del sujeto emocionado) y una función valorativa (que ayuda a la toma de decisiones). Así el reconocimiento de emociones a través de la voz podría ayudar a reducir la frustración de muchos humanos al interactuar con los sistemas automáticos y con los ayudantes basados en agentes inteligentes o incluso con juguetes, incrementando su empatía y simpatía (*R. Picard* 1997).

La voz sintética estándar suele responder a un modelo de voz neutra carente de matices emocionales, de actitud o de personalidad, lo cual le confiere características de monotonía y falta de naturalidad que no presenta la voz natural (*I. Murray* 1989). En la literatura se pueden encontrar diversos estudios sobre los correlatos vocales de los estados emocionales de los seres humanos (*I. Murray & J. Arnott* 1993). Aunque la mayoría de los efectos observados se refieren a la prosodia (velocidad de elocución, nivel, rango e inflexiones de F0 e intensidad), también se encuentran efectos en la articulación y la cualidad de voz (*J.M. Montero et al* 1998), (*N. Campbell* 2000).

Aunque las emociones estudiadas son muchas y variadas (*K. Scherer* 2000), 3 se encuentran recogidas en la práctica totalidad de los estudios y suelen estar entre las catalogadas como primarias, básicas o fundamentales[13]: felicidad, tristeza y enfado. Otras emociones estudiadas son: miedo, sorpresa, pena, repugnancia, vergüenza, etc.

2.4.1 Síntesis por formantes

Este método de síntesis intenta modelar matemáticamente la voz inspirándose en el modelo de producción de voz fuente-tracto-radiación propuesto por Fant e intentando modelar el filtro del tracto por medio de sus resonancias o

formantes (*D. Klatt* 1987). Para modelar cada resonancia se suele emplear un filtro sintonizado a la frecuencia deseada (variable en el tiempo). La combinación de los filtros para formar el modelo del tracto puede ser en cascada, en paralelo o mixta. Una rama serie es buena produciendo sonido sonoros no nasales, que sólo necesitan un control de amplitud, precisando un anti-resonador adicional para los sonidos nasales. Por su parte la rama paralela es mejor para sintetizar sonidos sordos y controlar la amplitud de cada resonador.

Este tipo de sistemas se suelen basar en reglas contextuales y tablas de formantes, siendo difícil su aprendizaje automático o su sustitución por métodos conexionistas (*M.S. Scordilis & Gowdy* 1990), debido a la dificultad para obtener datos de entrenamiento bien etiquetados, y a las limitaciones del modelo matemático en sí.

2.4.2 Sistemas de síntesis de voz con emociones

El diseño e implementación de sistemas de síntesis de voz con emociones comenzó a finales de los ochenta con el *Affect Editor* (*J. Cahn* 1989) y *Hamlet* (*I. Murray & J. Arnott* 1995).

2.4.2.1 El sistema Affect Editor

No es un sistema completo de conversión texto a voz, sino un editor de parámetros del modelo de Klatt empleado por *DECTALK*,. Experimenta con los parámetros prosódicos típicos (nivel medio de F0, pendiente de la curva de F0, rango de F0, comportamiento prepausa, velocidad media de elocución, etc.) y con algunos parámetros de fuente (ruido de fricación en sonidos no fricativos, brillantez o tilt, etc.) y un parámetro articulatorio (precisión de articulación). Las emociones evaluadas son: enfado, asco, agrado (glad), tristeza, miedo y sorpresa. La tasa de identificación está en torno al 50%, salvo para la tristeza que alcanza el 91% (28 sujetos, 5 frases y un párrafo de toma de contacto inicial).

2.4.2.2 El sintetizador Hamlet

Emplea igualmente el sintetizador de formantes *DECTALK* para simular 6 emociones (felicidad, tristeza, enfado, miedo, pena y repugnancia). Es un sistema de conversión texto a voz completo con 2 partes:

- **Proceso pre-acústico:** la conversión grafema a fonema de *DEC*, la prosodia (el modelo *Hat* de F0 y el multiplicativo de duración, procedentes de *MITTALK*[14]) y un nuevo sistema de reglas específicas para síntesis con emociones (que permite incrementar o decrementar los parámetros prosódicos y de cualidad de voz, de manera que las características emocionales son modificaciones sobre una voz neutra que sirve de base y que conservan las características básicas de acentuación o *breathiness* originales).
- **Proceso acústico**: el sintetizador *DECTALK* trabajando en modo fonema, para poder modificar su entonación y sus duraciones por defecto.

En *Hamlet*, para la configuración del sintetizador y la confección de las reglas dependientes de emoción, se han basado en la literatura previa y en heurísticos desarrollados por los autores según su percepción. Los parámetros empleados son: velocidad de elocución (*speech rate*), tono medio (*average pitch*), rango de tono (*pitch range*), pendiente de declinación, intensidad (*loudness*), *tilt* espectral, cuarto formante serie, etc. Las reglas son 11 y modifican parámetros fundamentalmente prosódicos (duración y F0 de cada fonema, tonema final y pausas), aunque también pueden cambiar la reducción de algunas vocales.

2.4.3 Prótesis vocales

Chat es un sistema de comunicación para personas con discapacidad vocal y motora, que incorpora predicción de expresiones para reducir el número de pulsaciones necesarias para escribir y sintetizar un mensaje. *Hamlet*, fue integrado en *Chat* y permite personalizar la voz al usuario, dotándola de una individualidad que es una característica muy apreciada por el colectivo de posibles usuarios (*I. Murray & J. Arnott* 1995).

En esta Tesis se describirá el desarrollo de un sistema de síntesis por formantes adaptable o personalizable para un usuario por parte de un experto.

2.4.4 Bases de datos de voz con emociones

La mayoría de estas bases no contienen voz grabada en condiciones emotivas naturales, sino provocadas o simuladas, aunque entre otros existen ejemplos de habla con estrés en (*J. Hansen & S. Bou-Ghazade* 1997), de situaciones muy limitadas como en (*J. Trouvain & W. Barry* 2000) o basadas en grabaciones televisivas o radiofónicas como en (*E. Douglas-Cowie, R. Cowie & M. Shröeder* 2000).

La mayoría de las bases de datos contienen textos neutros y emociones simuladas por parte de uno o varios actores (o bien locutores puestos en situación o estimulados por texto de contenido emocional), grabados en condiciones de estudio como (*I. Engberg, A. Hansen et al* 1997), o la que es objeto de la presente Tesis (*J.M. Montero et al* 1998). Es de observar que la utilización de actores puede ser buena para la labor comunicativa, pero puede dificultar la identificación de un usuario de una prótesis vocal con la voz sintetizada (*N. Campbell* 2000).

2.4.4.1 Bases de datos en castellano

Proyecto VAESS: es la desarrollada y descrita en la presente Tesis.

Proyecto INTERFACE: donde participa la UPC (*A. Nogueira et al* 2001), donde se ha grabado a un actor y a una actriz simulando las 6 emociones del estándar MPEG-4 (enfado, asco, miedo, alegría, tristeza y sorpresa, además del estilo neutro). Han grabado párrafos, frases y palabras aisladas.

En (*I. Iriondo* 2001) se describe una base de datos multilocutor (8 actores), multiemoción (alegría, deseo, furia, miedo, sorpresa, tristeza y asco) y multintensidad (3 niveles) evaluada con un gran número de oyentes (más de 1000), con la cual analizan diversos parámetros prosódicos generales (duración de las pausas, duración de los grupos fónicos, valor medio y rango de F0…) y un parámetros espectral (energía en bandas) para caracterizar las distintas simulaciones de voz con emociones. No llegan a realizar conversión texto habla automática (edición de la prosodia generada por un conversor).

2.4.5 Evaluación de sistemas de voz con emociones

Es muy difícil realizar una excelente evaluación de un sistema de voz sintética con emociones. Si una de las funciones de las emociones es transmitir información sobre el estado del sujeto y provocar reacciones en los demás, el contexto desempeñaría un papel no desdeñable, y difícil de reproducir. La experimentación de laboratorio y los entornos cotidianos son en gran medida incompatibles. Por eso la evaluación de los sistemas descritos (y otros descritos en la bibliografía) se basa en experimentos de laboratorio de identificación de emociones simuladas o de opinión sobre la calidad de la voz emotiva.

Los contenidos de los textos de evaluación han sido objeto de investigación, pues existen 3 tipos de textos: con contenido emotivo, con contenido neutro y sin contenido emotivo definido. Las respuestas de los sujetos pueden ser igualmente de 3 tipos: respuesta forzada, respuesta libre y pares opuestos.

La evaluación de (*J. Cahn* 1989) es de respuesta forzada (con un campo de respuesta abierta) y textos de contenido neutro, mientras que en (*I. Murray & J. Arnott* 1995) se combinan la respuesta libre y la respuesta forzada, los textos con contenido emotivo y los textos neutros[15]. La evaluación se suele realizar con auriculares, en salas tranquilas.

Los resultados son bastante variables y no siempre de fácil interpretación. El sistema Hamlet evidencia una baja tasa de reconocimiento de la voz neutra como neutra en un experimento sólo con voz neutra, que sólo mejora considerablemente cuando se pone, en el mismo *test*, voz simulando emociones con textos emotivos. La felicidad, el miedo y la pena son percibidos mayoritariamente como voz sin emociones si la voz es neutra pero el contenido no. Cuando el contenido es neutro y la voz simula emociones, predominan las respuestas de tristeza, que es junto al enfado la única emoción mayoritariamente identificada. Cuando la voz es emotiva y el texto apropiado a dicha emoción, la matriz de confusión entre emoción simulada y emoción percibida se vuelve, por fin, casi diagonal. Experimentos como los realizados en esta Tesis parecen indicarnos que o bien algunas emociones no eran correctamente sintetizadas, o bien su percepción depende de parámetros no modelados por estar poco relacionados con la prosodia (*J.M. Montero et al* 1998), pues la falta de experimentos con voz natural nos impide sacar conclusiones sobre la identificabilidad intrínseca de las emociones empleadas. Sí que podemos concluir que el contenido de los textos puede influir significativamente sobre la identificación de emociones simuladas, y que conviene emplear textos neutros para evaluar la capacidad de la voz sintética para simular una emoción.

Capítulo 3 Procesado lingüístico automático

3.1 Introducción

Como ya hemos mencionado, Boris, el sintetizador del GTH, dispone de un módulo de procesamiento lingüístico orientado a prosodia. Su diseño robusto y minimalista vino influido por el objetivo de minimizar el consumo de recursos del sistema medido en términos de memoria RAM utilizada, buscando un sistema global de muy bajo coste. No empleaba diccionarios (aunque sí listas de excepciones) porque buscaba minimizar la memoria utilizada, pensando en el desarrollo de sistemas muy baratos en un ordenador personal.

El paso del tiempo y la evolución de la informática doméstica ha hecho abordable el problema por la mayor disponibilidad de recursos informáticos y en esta Tesis analizaremos cómo emplear el conocimiento lingüístico previamente existente en el GTH, cómo integrarlo con nuevas fuentes léxicas y cómo mejorar el sistema global empleando información probabilística.

El sistema de categorización por regla del GTH empleaba un conjunto de etiquetas distinto e independiente del conjunto definido en el corpus 860. El hecho de pasar a emplear un diccionario hace que muchas de las reglas desarrolladas no tengan ahora la efectividad para la que fueron pensadas. La detección de verbos se ve considerablemente mejorada cuando se incorpora un módulo de conjugación y una amplia lista de infinitivos (tanto regulares como irregulares). La distinción entre sustantivos y adjetivos no se tenía en cuenta, pudiendo tener importancia a la hora de desambiguar las categorías de otras palabras, aunque posiblemente tenga poca a la hora de analizar sintácticamente.

Las reglas originales respondían a un formato con capacidades de transformación, aunque usualmente se encargaban de categorizar palabras que previamente no habían podido ser etiquetadas. Frente a este planteamiento, en la experimentación con reglas se tenderá a emplear reglas de selección, que partiendo de una categorización ambigua seleccionen o eliminen posibilidades.

3.2 Etiquetado morfosintáctico automático

3.2.1 Corpora empleados

3.2.1.1 El corpus de El Mundo

A fin de probar el funcionamiento de los mecanismos de segmentación en frases y de detección de palabras especiales emplearemos el conjunto de textos del periódico El Mundo correspondiente a los años 1994, 1995 y 1996, disponibles en CD-ROM. De acuerdo con nuestros cálculos, el corpus de entrenamiento (1994 + 1995)

contiene 54.012.863 palabras, 2.120.328 frases, distribuidas en 96.529 textos. Para evaluación se han empleado varios conjuntos de textos del año 1996, como se mencionará más adelante.

3.2.1.2 El corpus 860

Está constituido por textos etiquetados manualmente durante el proyecto ESPRIT 860, que se pueden clasificar en tres tipos:

1) Textos jurídicos y de legislación (**EEC**), que constituyen un 30 % del total.
2) Textos de la Comunidad Europea (**CEE**), un 63.08 %.
3) Textos periodísticos (**TEXTSPA**), un 6.92 %.

Número de ...	Etiquetas completas	Etiquetas simplificadas
elementos del conjunto de evaluación del 860	38.172	38.172
etiquetas diferentes presentes en el corpus	345	236
palabras fuera de vocabulario (*OOV*)	1.244 (3,2589%)	1.244 (3,2589%)
palabras con etiquetas fuera de diccionario, que no son *OOV*	78 (0,2043%)	71 (0,186%)
tipos de ambigüedades (*ambiguity classes*)	911	875
etiquetas por palabra en el corpus	2,9043	2,1694
palabras ambiguas en el conjunto de evaluación	261	252
palabras ambiguas en el corpus	911	875

Tabla 1 Principales parámetros del corpus 860

Número de ...	Etiquetas completas	Etiquetas simplificadas
elementos del conjunto de entrenamiento del 860	278.378	278.378
elementos diferentes en el corpus	20.857	20.857
elementos diferentes en el conjunto de entrenamiento	19.677	19.677
elementos diferentes en el conjunto de evaluación	6.625	6.625
unigramas y etiquetas diferentes en el conjunto de entrenamiento	331	227
unigramas y etiquetas diferentes en el conjunto de evaluación	270	190
bigramas diferentes en el conjunto de entrenamiento	8.194	5.388
trigramas diferentes en el conjunto de entrenamiento	40.576	30.910
bigramas diferentes en el conjunto de evaluación	3.599	2.568
trigramas diferentes en el conjunto de evaluación	12.032	10.093
etiquetas fuera de vocabulario	47	31

Tabla 2 Parámetros secundarios del corpus 860

Existe un cuarto tipo de textos relacionados con la temática de rehabilitación, discapacidad y ayudas técnicas, que se etiquetó en un proyecto posterior y contiene unas 52.065 palabras y signos de puntuación. El conjunto de etiquetas empleados puede ser consultado en el apéndice A.1.1 “Etiquetado del 860”, junto con una revisión crítica del mismo.

3.2.1.2.1 Parámetros principales del corpus 860

Siguiendo las recomendaciones de EAGLES sobre evaluación de etiquetado automático (*G. Leech et al 1*998), podemos caracterizar nuestro corpus según las tablas “Principales parámetros del corpus 860” y “Parámetros secundarios del corpus 860”

De acuerdo con las categorías primarias empleadas en el etiquetado, la distribución de probabilidad de las mismas se puede ver en la siguiente figura:

Tabla 3 Comparación entre la distribución de los distintos símbolos (en tanto por uno) en los corpora de entrenamiento, de evaluación y el completo: Verbo, Nombre sustantivo, Adjetivo, adverBio, pRonombre, Preposición, Determinante, Conjunción, Interfección, Miscelanea y otros (L)

Donde el coeficiente de la correlación de Pearson entre las distribuciones de entrenamiento y evaluación es muy alta (**0,99989**), dado el carácter aleatorio de la partición.

3.2.1.2.2 Problemas detectados en el etiquetado manual del corpus del 860

Como consecuencia del trabajo de cobertura léxica realizado, y frecuentemente debido a emplear fuentes léxicas distintas del propio corpus 860 -tales como el DRAE u Onomástica (*The Onomastica Consortium* 1995)-, se ha detectado la existencia de diversos problemas en el corpus que hemos considerado necesario revisar, corregir o adaptar. Como ya hemos señalado previamente, todo etiquetado manual conlleva inevitablemente errores, motivados por falta temporal de concentración, por la diversidad de etiquetadores implicados (aunque en todo caso se trate de expertos), por la dificultad para seguir un conjunto de normas (quizá no suficientemente interiorizadas), etc.

3.2.1.2.3 Errores triviales y de selección

§ **Falsos verbos**: la *parte* correspondiente, las *estructuras* industriales, la *IBA* británica, la *estructura* sectorial, tanto *fijas*, de *ayuda*...

§ **Falsos sustantivos**: orientaciones *propuestas*, canal *multinacional*, iniciativas *originales*, lo *posible*, que *toque*, las *grandes*, se *evidencia*, el *pasado* ejercicio, no *cuenta*, les *cuesta*, *cuesta* saber, *once*, quedó *patente*, los *demás*, productos *textiles*, materias *primas*, estrategia *industrial*, marinos *griegos*...

§ **Falsos adjetivos**: las *corrientes* de, se *adjunta*, *cierto* (adjetivo calificativo en vez de indefinido) tiempo, *diario* oficial...

§ **Falsos adverbios**: *tal*+sustantivo (naturaleza, acuciamiento, guerra, magnitud)...

§ **Falsos pronombres**: la *NOS*, sí *mismo*, ellos *mismos*, [lo, la, los, las] + **N**, [lo, la, los, las] + **A**, [lo, la, los, las] + **P**, [lo, la, los, las] + **M**, *las* indicadas, *la* impugnada...

§ **Errores en el pronombre que**: lo *que*, los *que*, los créditos *que*, beneficiarios *que* presenten, la primera vez *que*...

§ **Otros errores en las categorías principal o secundaria**: *extra* (P*), *del* (P00*)...

§ **Errores en rasgos secundarios**: tiempo verbal (*incumbe*. V..03*), enclíticos (*cállate*. V*##..), número (*mero* *P*, *marginado* *P*), género (*poemas* *F*, *facilidad* *M*)...

§ **Etiquetado semántico en vez de morfosintáctico**: *propio* (etiquetado como posesivo), *nuestro* y *nuestros* (comparten igual rasgo persona, en vez de contener diferente rasgo número).

3.2.1.2.4 Diferencias e irregularidades de criterio

Aunque los errores en los determinantes son también irregularidades o diferencias entre etiquetadores, en este apartado debemos destacar la falta de uniformidad en el etiquetado de las **locuciones** o unidades léxicas multipalabra (*en concreto, en absoluto, de inmediato, al contrario, al por menor, frente al, al frente de, al mismo tiempo, por medio de, de en medio, después de, detrás de, así como, aparte de, sin embargo, de que, alrededor de, salvo en*...) o incompletitud (por ejemplo, se contemplaba la locución *con motivo de*, pero no la locución *con motivo de que*).

3.2.1.2.5 Diferencias de formato

Dado que nuestro interés por el etiquetado viene motivado por nuestro interés en la síntesis de voz, deseábamos tener un módulo de procesamiento de texto sin restricciones, aunque los experimentos de etiquetado los realicemos finalmente sobre texto pre-procesado.

§ **Tratamiento de signos intra-palabra**: belgo-luxemburguesa, extra-comunitarios, Euro-show,

socio-económica, socio-comunitarias, nacional/internacional, opto-electrónica, histórico-artísticas, histórico-artístico.

§ **Puntos suspensivos en líneas separadas**.

§ **Errores del texto**: *per juicio, le* ...(N).

§ **Formato de cifras**: empleo irregular de comas o puntos.

§ **Números romanos**: en minúsculas.

§ **Siglas**: divididas en una secuencia de letras**.**

§ **Mezcla de letras y números**: separados.

La imprecisión global detectada en los textos del 860 en el corpus de evaluación (38172 elementos léxicos) ha sido del **1,8862%,** aunque si incluimos las diferencias de criterio habría que añadir un **1,7604%** más.

3.2.2 Modelado léxico

Para el etiquetado inicial hemos empleado un procesamiento basado en una secuencia de 4 tipos de procesadores: normalizador, modulo de diccionarios, conjugador verbal y reglas de terminaciones.

3.2.2.1 Normalizador

Aunque el corpus que emplearemos para evaluar el etiquetado automático ha sido normalizado manualmente dentro del marco del proyecto 860, el sistema de etiquetado que se ha desarrollado contiene un módulo de normalización que ha sido evaluado con un corpus diferente, como ya se mostró en (*A. Jiménez* 1999).

Para la segmentación en frases, por medio de **reglas léxicas internas** al programa de etiquetado, se detectan las palabras que son terminadores de oración (punto, fin de párrafo, signos de interrogación o admiración) o que los contienen:

1) **Fechas**: se admite el formato en el que el separador es un punto, y no sólo aquellos donde el separador es un guión o una barra inclinada. Se admite una mezcla de números y nombres de meses.

2) **Horas**: al exportar el texto del periódico se pierde parte del formato, dando lugar a horas precedidas por una K y con un punto intermedio.

3) **Expresiones numéricas**: no exigimos un formato castellano estándar, admitiendo cualquier combinación con comas y puntos.

4) **Códigos alfanuméricos**.

5) **Abreviaturas y símbolos**: admitimos que comiencen por mayúscula o minúscula y que tengan puntos intermedios o sólo punto final, y empleamos un diccionario auxiliar que contempla las más comunes, obtenidas a partir de listas de la RAE y la ISO.

6) **Siglas**: para ser detectadas deben estar escritas en mayúsculas con o sin puntos intermedios. También en este caso se emplea un diccionario auxiliar. Es también importante eliminar los títulos y las firmas de los artículos dado que suelen ir en mayúsculas. Limitándonos a detectar siglas que contengan entre 3 y 5 letras, el error cometido es un 2,23 %.

7) **Puntos suspensivos**: debido a errores tipográficos debemos aceptar no sólo el modelo estándar que contiene 3 puntos sino también un formato con sólo 2.

8) **Palabras con guión o barra inclinada**: esta categoría incluye los guiones que se utilizan para expresar género o número (**señor/a**) y los guiones o barras que se utilizan para la formación de compuestos: sustantivo-sustantivo como *ciudad-dormitorio*, adjetivo-adjetivo como *histórico-artísticos*, verbo-verbo como *espulga/expurga*, adverbio-adverbio como *arriba/abajo*, prefijo-sustantivo, prefijo-adjetivo o prefijo-verbo como en *ante-sala* o en *sobre-dosis*.

Evaluado sobre un corpus no visto durante el desarrollo de las reglas, la segmentación en frases tuvo una tasa de

error de tan sólo un 0,42% (*A. Jiménez* 1999) sobre un pequeño conjunto de 479 frases.

Otros ejemplos de palabras especiales son:

1) **Números romanos**: donde se comprueban formatos no válidos para evitar tomar algunas siglas como números romanos. Se debe analizar la palabras anterior y la siguiente a fin de no confundirlos con las siglas.

2) **Nombres propios**: cualquier palabra que comience por una mayúscula seguida por minúsculas en posición no inicial dentro de una frase es un serio candidato a ser un nombre propio. Si se encuentra en posición inicial, debemos consultar los varios diccionarios de nombres propios que se comentan más adelante (el proveniente del 860, uno de nombres de pila de ONOMASTICA y otros dos de apellidos y poblaciones de la misma procedencia). Un caso especial lo constituyen los nombres propios de más de una palabra: secuencias que identifican a personas, entidades, etc., que admite abreviaturas intermedias, guiones o las preposiciones *de* y *del*. De esta manera, sobre un conjunto de 10.544 frases, la precisión fue del 96,22 %, donde la mayor parte del error se debió a la falta de cobertura en posición inicial.

4) **Combinaciones de letras, números y signos.**

5) **Palabras compuestas con guiones**.

6) **Palabras extranjeras con caracteres especiales.**

7) **Letras aisladas.**

La detección de Números, Números con guión, Fechas, Horas y Abreviaturas obtienen una precisión del 100 % (todas las unidades detectadas pertenecen a la categoría asignada).

Unidad especial	Total	Aciertos	Error (%)
Siglas de diccionario	2.087	1.801	11,58
Otras siglas	6.965	6.881	2,23
Nombres Propios	5.957	5.891	1,11
Nombres Propios dudosos	4.867	4.831	0,74
Nombres Propios compuestos registrados en el diccionario del 860	1.690	1.652	2,25
Nombres Propios compuestos de diccionario	2.121	1.939	8,58
Nombres Propios compuestos dudosos	1.558	1.529	1,80
Nombres Propios compuestos con las palabras 'de' o 'del'	2.115	1.979	6,43
Nombres Propios compuestos con abreviatura	6.868	6.124	10,83
Números romanos	2.102	1.850	11,99
Palabras con guiones	2.037	1.801	11,58

Tabla 4 Resultados de imprecisión para varios tipos de palabra no normalizadas, según la información léxica empleada en su detección

El contraste entre estas cifras y las generales que daremos posteriormente sobre el corpus del 860, reflejan la dificultad que entrañan estas palabras especiales dado que reflejan hechos lingüísticos de notable creatividad, y que el dominio sobre el que se desarrolle un sistema influye especialmente sobre este tipo de unidades (a pesar de que tanto el corpus de El Mundo como el del 860 contemplan textos de periódico).

3.2.2.1.1 Detección de palabras extranjeras o mal escritas

Durante el desarrollo del sintetizador se implementó un silabicador basado en la detección de aquellos pares de letras que no se pueden dar dentro de una sílaba castellana.

Este módulo de silabicación es la base de un método muy simple para la detección de algunas palabras extranjeras o mal escritas. Dada una palabra que no esté presente en los diccionarios, que esté escrita en minúsculas y que no haya sido reconocida por el conjugador verbal, si como resultado de la división en sílabas, detectamos la presencia

de alguna sílaba sin vocales o que contenga un par de consonantes repetidas consecutivas no posibles en castellano (o sea, todas aquellas que no sean *l, r, c*) y las clasificaremos como extranjeras o mal escritas. Evaluadas sobre 128 palabras extranjeras y 1.129 nombres propios extranjeros detectados, las tasas de imprecisión fueron **0,78 %** y **0,62 %**, respectivamente.

3.2.2.2 Diccionarios

En el sistema de etiquetado disponemos de varios diccionarios etiquetados:

§ **Las palabras más frecuentes del proyecto 860**: con unas 15.000 palabras, revisadas manualmente, aunque en los experimentos de desambiguación se empleará sólo una parte del mismo.

§ **El DRAE**: con unas 150.000 palabras, aunque no es tan fiable como el anterior en cuanto a su clasificación de las palabras función.

§ **Los nombres, apellidos y poblaciones**: provenientes del proyecto europeo Onomástica.

§ **Los diccionarios de inglés, francés e italiano del proyecto MOBY**: el coste computacional que supone su empleo sólo compensa para el caso del inglés., o para un corpus donde sean frecuentes las palabras extranjeras. Sobre los textos de El Mundo: diccionario *English* de 182.790 palabras: 5.124 encontradas (2,80%); diccionario *French* con 132.418 palabras: 996 encontradas (0,75%); diccionario *German* con 159.587 palabras y 1.077 encontradas (0,67%); diccionario *Italian* con 60.453 palabras y 450 encontradas (0,74%). Para obtener estos resultados fue necesario eliminar expresiones en castellano del diccionario de inglés (*A. Jiménez* 1999) y evaluar sobre un nuevo conjunto de evaluación.

§ **Los diccionarios de locuciones del 860**: convenientemente completados de tal manera que si se incluye la locución *por encima de*, también se incluya *por encima del* y viceversa o si se incluye la locución *debido a*, se incluya también *debido al*, aunque no hayan sido registradas en el conjunto que sirva de entrenamiento (excepciones pueden ser con *el fin de* o *en caso de*, aunque esta sobregeneración no conducirá a error). Otra posibilidad de ampliación se basa en la complementariedad entre las locuciones que terminan en la palabra *de* y las que terminan en *de que*, como *a pesar de* y *a pesar de que*, aunque este mecanismo introduce más sobregeneración (*a nivel de* es una locución, pero *a nivel de que* no, puesto que la primera no rige verbos en infinitivo y es puramente prepositiva).

§ **Los diccionarios de infinitivos regulares e irregulares de la RAE**: dado que disponemos de un conjugador verbal sólo se necesita tener unos diccionario de infinitivos para que el sistema pueda detectar formas verbales. Para evitar errores frecuentes en la detección de verbos, se han eliminado infinitivos de poco o nulo uso en la actualidad.

La búsqueda se realiza de manera binaria apoyada en índices, aunque se podrían emplear técnicas más rápidas como las basadas en una tabla *Hash* o en un *letter-tree index* (*trie*).

3.2.2.3 Conjugador verbal

Para el tratamiento de los verbos, dada la compleja conjugación del castellano, ha sido necesario desarrollar un completo conjugador verbal. Dentro de la jerarquía de *Hockett* (*A .L. González et al* 1995), hemos optado por el modelo más simple, el de Palabras y Paradigmas. Las palabras (en este caso los verbos) se clasifican en paradigmas, regulares o irregulares, caracterizados por una raíz y una terminación que se deben concatenar para crear una forma verbal. Este tipo de procesamiento permite implementaciones muy flexibles y eficientes basadas en autómatas finitos, aunque no ha sido objeto de la presente Tesis. Los paradigmas irregulares pueden consultarse en el apéndice A.1.2 "*Lista de paradigmas irregulares empleados"*. No se han incorporado todos los verbos defectivos, pudiendo producirse un fenómeno de sobregeneración.

3.2.2.4 Reglas léxicas externas o de terminaciones

Finalmente, para las palabras fuera de vocabulario es necesario incorporar reglas de terminaciones. Para ello partimos de las reglas disponibles en el grupo (*J. A. Vallejo* 1998), de las desarrolladas para el castellano en el

proyecto CRATER (*F. Sánchez-León & A. Nieto* 1997), aunque las diferentes condiciones de aplicación (ausencia de diccionario o menor cobertura de este), pueden hacerlas mucho menos útiles.

Es importante señalar que las palabras de morfología más irregular son las más frecuentes, y que las palabras nuevas (no incluidas en los diccionarios) suelen seguir los paradigmas más regulares.

Las reglas externas previas estaban diseñadas para un módulo sin diccionarios previos; la incorporación de mucha información léxica hizo necesario depurar las existentes, eliminando las reglas verbales y algunas relativas a palabras extranjeras (muy poco útiles ahora que contamos con un conjugador). De esta manera se aumentó su precisión de un 77,5% a un 98,88% (es de observar que, en general, las reglas respondían a fenómenos lingüísticos generales que podían alcanzar gran precisión). Partiendo de un conjunto de 117 reglas nos quedamos con 77 (estas reglas pasarían de categorizar un 31,8% de las palabras que le llegaban, a etiquetar un 24,8%).

De manera similar y por los mismos motivos, filtramos las reglas del proyecto CRATER, incrementando la precisión de un 88,9% a un 96,81%. Partiendo de un conjunto de 217 reglas que etiquetaban el 98,7% de las palabras, nos quedamos con 148 reglas cuya cobertura es del 54,3%.

Finalmente se añadieron 162 nuevas reglas de poca cobertura (15%) pero gran precisión (98,85%).

3.2.2.5 Cobertura léxica

En el caso del corpus 860 gran parte de nuestro sistema léxico debe ser dividido a fin de disponer de unos conjuntos independientes de entrenamiento y evaluación, aunque la disponibilidad de recursos lingüísticos adicionales va a minimizar la falta de cobertura.

Diccionarios de...	Número de palabras	Número de pares palabra-categoría	Ambigüedad
Corpus de entrenamiento del 860	19.676	20.689	1,0513
Corpus de evaluación del 860	6.625	6.971	1,0522
Infinitivos regulares	10.087	10.087	1
Infinitivos irregulares	2.589	2.589	1
Real Academia Española	160.314	160.355	1,00026
Locuciones	263	263	1
Siglas	216	216	1
Nombres Propios del 860	2.050	2.050	1
Nombres de pila de Onomástica	8.706	8.706	1
Apellidos sencillos de Onomástica	49.207	49.207	1

Tabla 5 Resumen de los diccionarios que se emplean en el modelado léxico

Partiendo del sistema que desarrollamos en (*A. Jiménez* 1999), se reemplazarán los diccionarios *860* por diccionarios de entrenamiento, a fin de separar los conjuntos de entrenamiento y evaluación. Esto afecta tanto al diccionario principal, como a los diccionarios de siglas, de nombres propios y de locuciones. Debido a la incompletitud e incoherencia del etiquetado de las locuciones, ha sido necesario homogeneizar el etiquetado de las locuciones. De esta manera se obtienen los resultados de la Tabla 6:

Debido al especial tratamiento de las palabras especiales que realizan las **reglas codificadas internamente** dentro del programa (números romanos, letras, fechas, horas, palabras con signos, números, siglas, nombres que comienzan por mayúscula y abreviaturas), la cobertura de un subconjunto del entrenamiento (entr_7) no es total. Sin embargo, este tratamiento permite una mejor cobertura de textos nuevos que se empleen para evaluación, con formato sin duda diferente al del proyecto 860.

En el conjunto de etiquetas que denominamos simplificadas se ha eliminado la distinción entre verbos transitivos, intransitivos y verbos que rigen la conjunción *que* (la transitividad e intransitividad se trata de un fenómeno más de estructura que de léxico, y cuyo ámbito sintáctico no es local). Aquellas palabras que pueden ser masculinas y femeninas se etiquetan como neutras. Tampoco se distingue entre palabras con género o número neutros,

invariantes o desconocidos. Sólo se distingue entre verbos con y sin enclíticos (su presencia elimina la posibilidad de que pronombres personales objetos precedan al verbo). No se distingue entre nombre comunes, propios y siglas, ni se distinguen varios tipos de expresiones numéricas.

Dado que el empleo de etiquetas simplificadas tiende a ofrecer resultados ligeramente mejores, y esta diferencia resulta significativa cuando se emplean todos los recursos disponibles (0.9996>0.9987), en esta Tesis ofreceremos resultados para el conjunto de etiquetas simplificadas.

Etiquetas	Diccionarios empleados	Reglas léxicas (internas y/o externas)	Fichero de evalua-ción	Número de palabras	Cober-tura	Número de etiquetas por palabra	Intervalo de confianza al 99%
Completas	Todos	Todas	entr_7	40.901	0,9993	1,6990	0,0003
Completas	Todos	Todas	Eval	38.172	0,9987	1,6984	0,0005
Simplific.	Todos	Todas	entr_7	40.901	0,9998	1,6146	0,0002
Simplific.	Todos	Todas	Eval	38.172	0,9996	1,6128	0,0003
Completas	Sin DRAE	Todas	entr_7	40.901	0,9993	1,6758	0,0003
Completas	Sin DRAE	Todas	eval	38.172	0,9969	1,6756	0,0008
Simplific.	Sin DRAE	Todas	entr_7	40.901	0,9998	1,5961	0,0002
Simplific.	Sin DRAE	Todas	eval	38.172	0,9975	1,5947	0,0007
Completas	Todos	Internas	entr_7	40.901	0,9993	1,5927	0,0003
Completas	Todos	Internas	eval	38.172	0,9966	1,5867	0,0008
Simplific.	Todos	Internas	entr_7	40.901	0,9997	1,5111	0,0002
Simplific.	Todos	Internas	eval	38.172	0,9972	1,5035	0,0007
Completas	Sin DRAE	Internas	entr_7	40.901	0,9993	1,5570	0,0003
Completas	Sin DRAE	Internas	eval	38.172	0,9924	1,5491	0,0012
Simplific.	Sin DRAE	Internas	entr_7	40.901	0,9997	1,4790	0,0002
Simplific.	Sin DRAE	Internas	eval	38.172	0,9930	1,4696	0,0011
Completas	Sin entrenam.	Todas	entr_7	40.901	0,9713	1,6114	0,0021
Completas	Sin entrenam.	Todas	eval	38.172	0,9706	1,6093	0,0022
Simplific.	Sin entrenam.	Todas	entr_7	40.901	0,9718	1,5194	0,0021
Simplific.	Sin entrenam.	Todas	eval	38.172	0,9712	1,5258	0,0022
Completas	Sin entrenam.	Internas	entr_7	40.901	0,9445	1,4334	0,0029
Completas	Sin entrenam.	Internas	eval	38.172	0,9438	1,4362	0,0031
Simplific.	Sin entrenam.	Internas	entr_7	40.901	0,9450	1,3547	0,0029
Simplific.	Sin entrenam.	Internas	eval	38.172	0,9444	1,3563	0,0030

Tabla 6 Resultados de etiquetado automático sin desambiguación contextual

Como era lógico la eliminación del diccionario de apoyo DRAE o de las reglas de terminaciones afecta muy poco al resultado sobre el subconjunto de entrenamiento mientras usemos el diccionario de entrenamiento que recoge las palabras más habituales en este domino, pero resulta significativa con el 99% de confianza cuando experimentamos con el conjunto de evaluación (0,9969 < 0,9987, 0,9975 < 0,9996).

La eliminación del diccionario de entrenamiento (aún conservando las palabras función y los verbos, y eliminando tan solo los sustantivos, adjetivos y adverbios) provoca una caída muy significativa de la cobertura, aún cuando se use el diccionario de apoyo (0,9706 < 0,9987; 0,9712< 0,9996) y, especialmente, cuando se suprimen las reglas de terminaciones (0,9433 < 0,9706; 0,9444<0,9712). En este caso no existen diferencias significativas entre los subconjuntos de entrenamiento y evaluación (0,9713 frente a 0,9706; 0,9718 frente a 0,9712; 0,9445 frente a 0,9438; 0,9450 frente a 0,9444). Destaca la aparición de nuevos errores debidos a formas verbales y adjetivales que no se contemplan como sustantivos en el entrenamiento del 860, pero sí en el DRAE.

Si no se emplean las reglas léxicas externas o de terminaciones, y tan sólo se emplean los diccionarios (antes

enumerados) y las reglas internas (las de palabras especiales), la tasa cae significativamente (0,9966 < 0,9987, 0,9972 < 0,9996). De nuevo el especial tratamiento de las palabras especiales hace que la cobertura del conjunto de entrenamiento no sea total, aunque superior a los resultados obtenidos por los métodos de aprendizaje automático antes reseñados. Los nuevos errores se deben obviamente a las reglas suprimidas:

- Adjetivos funcionando como sustantivos: tales como *marinos* o *precedentes*.
- Nuevos adverbios en –mente: como *ocasionalmente* o *transitoriamente*.
- Otras palabras fuera de vocabulario de morfología conocida: como *sindicalistas*, *opto-electrónica* o *fotovoltaicas*.

Si sobre los subconjuntos de evaluación y en los mejores experimentos, analizamos los errores de cobertura, podemos clasificarlos como:

- **Locuciones nuevas**: como *cuanto que*.
- **Gentilicios:** que sólo están contemplados como sustantivos y no como adjetivos, como *irlandeses*.
- **Puntos cardinales:** también contemplados como sustantivos, como *sudeste*.
- **Otros errores adjetivo / sustantivo / participio:** así como en castellano la mayoría de los adjetivos calificativos pueden ser nominalizados por un determinante (y por ello se ha establecido una regla que añada la categoría sustantivo cuando se encuentre la categoría adjetivo calificativo), el caso complementario no está contemplado y produce algunos errores marginales.
- **Indefinidos:** incluidos como adjetivos pero no como adverbios, como *mucho*.
- **Nombres propios:** como *el programa quinquenal Valoren*.
- **Números romanos o letras:** como *ANEXO II* o *Recomendación X* o letras aisladas entre puntos.
- **Palabras extranjeras:** pero formadas por sílabas típicamente castellanas como *dello*.
- **Palabras fuera de vocabulario**: debidas al tipo de dominios empleados, como *os* o *conmigo*.

Varios de estos problemas (gentilicios, puntos cardinales e indefinidos) son solucionables de un modo fácil y general por medio de reglas o ampliando manualmente los diccionarios.

3.2.2.5.1 Un etiquetador léxico estocástico

Ficheros	Etiquetado	Cobertura	Etiquetas / palabra	Palabras desconocidas
Eval	Completo	0,9983	3,2891	0,0325
Eval	Simplificado	0,9983	2,5199	0,0325

Tabla 7 Resultados de cobertura léxica del etiquetador automático TnT

El etiquetador estocástico TnT (*T. Brants 1999*) emplea exclusivamente probabilidades para etiquetar; para la desambiguación contextual emplea trigramas, pero también contiene una parte léxica compuesta por un diccionario y un *suffix trie* probabilístico.

Empleando la parte léxica (con longitud de sufijo menor o igual que 10 y sin poda de caminos o *beam-search*), se obtienen resultados significativamente peores que nuestra mejor combinación anterior para etiquetas simplificadas (0,9983 < 0,9996), aunque dentro de una tasa de cobertura elevada. En el caso de etiquetas completas los resultados también son peores pero no significativamente.

3.2.3 Desambiguación contextual

Tras los módulos léxicos previos, el etiquetado de las palabras es ambiguo (incluso muy ambiguo). Es necesario poder aplicar nuevas técnicas que eliminen (o al menos reduzcan) esta ambigüedad, empleando para ello información acerca del entorno de cada palabra en cuestión.

3.2.3.1 Creación de reglas manuales contextuales

Tras los buenos resultados de las reglas léxicas externas, se decidió continuar con la tradición de reglas manuales del GTH, y se procedió a completar el conjunto de reglas disponibles. Las 102 reglas contextuales heredadas fueron evaluadas con malos resultados: la cobertura sobre el corpus de evaluación se reducía a 0,9360 (desde 0,9987) debido a que su precisión era sólo de 0,5930 (etiquetaban erróneamente muchas palabras). Reduciendo las reglas de 102 a 84 se consiguió subir la cobertura a un 0,9647. Rediseñando las reglas como reglas de selección (en vez de como reglas de transformación) se alcanzó una cobertura de 0,9787 con una ambigüedad reducida hasta 1,26 etiquetas por palabra.

Las reglas eran razonablemente fiables, pero su tasa de ambigüedad continuaba siendo elevada. Al aplicar un filtrado posterior que eliminó los pares de categorías no vistas en el conjunto de entrenamiento (los ceros de la *bigram*), se redujo la ambigüedad a 1,1413, pero reduciendo la cobertura al 0,9733. Las ambigüedades remanentes se pueden clasificar como:

- **Adjetivo-sustantivo**: el 32,96% de las ambigüedades en casos de sustantivos o adjetivos que no sabíamos discernir si eran una cosa o la otra.
- **Posesivos**: que pueden ser adjetivos determinantes o pronombres (1,31%).
- **Sustantivo o verbo**: el 29,68% de las ambigüedades en casos de sustantivos o verbos que no sabíamos discernir si eran una cosa o la otra.
- **Las palabras 'la' y 'las'**: pueden ser, fundamentalmente, artículos o pronombres (14,02%)
- **La palabra 'que'**: puede ser pronombre relativo o conjunción (15,33%).

Los errores cometidos por categorías son:

- **Sustantivos**: 641 errores.
- **Adjetivos**: 182 errores.
- **Adverbios**: 105 errores.
- **Verbos**: 46 errores.
- **Pronombres**: 44 errores.
- **Otros**: 38 errores.

Como el proceso manual de creación de reglas resultaba muy costoso, se procedió a emplear 2 técnicas alternativas: aprendizaje automático y modelado estocástico.

3.2.3.2 Aprendizaje automático de reglas

Siguiendo el paradigma transformacional guiado por errores desarrollado por E. Brill, procedimos a adaptar su sistema de dominio público en investigación para trabajar con nuestro corpus 860 etiquetado manualmente para castellano, dado que no se basa en un modelo especialmente dependiente del lenguaje.

Inicialmente se asigna a cada palabra (conocida o no) su categoría más probable en función del corpus de entrenamiento o de cualquier otra fuente, empleando un diccionario, reglas morfológicas o de terminaciones, otro categorizador previo, etc. A partir de ahí se aprende cómo corregir cada error cometido por medio de reglas contextuales de transformación, en un proceso voraz (*greedy*) iterativo y costoso de mejora. Para el aprendizaje de las reglas se parte de un conjunto de condiciones patrón (o *templates*) que permiten generar cada posible regla contextual o léxica concreta y probar su capacidad de corrección de errores por transformación.

Una vez realizada la adaptación de formatos y la selección de patrones, se puede proceder a entrenar el sistema. Se parte de la división de 860 en 3 dominios (periodísticos, legislación y Comunidad Europea).

3.2.3.2.1 Conjunto de etiquetas

Como el número de categorías del 860 es elevada, aplicamos una simplificación inicial que redujo su número, que

se puede consultar en el apéndice A.1.4 (“Conjuntos de etiquetas del experimento de aprendizaje de reglas de categorización”). Tras unos primeros experimentos se procedió a usar un conjunto más detallado a fin de mejorar los resultados, añadiéndose las categorías que se pueden ver en dicho apéndice.

3.2.3.2.2 Evaluación

A fin de probar la capacidad del aprendizaje para generalizar, se realizaron experimentos dentro y fuera de dominio, siempre manteniendo la separación entre conjunto de entrenamiento y conjunto de evaluación, manteniéndose entre ellos una *ratio* aproximada de 8 a 1; para el aprendizaje de las reglas léxicas se dividió el entrenamiento en dos partes de igual tamaño.

Los resultados se muestran en la siguiente tabla:

Entrenamiento	Tagset	Evaluación	Número de Palabras	Acierto	Intervalo de confianza al 95%
Dominio1	Inicial	Dominio 1	13.272	99,25 %	0,146
Dominio 2	Inicial	Dominio 2	16.530	99,50 %	0,107
Dominio 3	Inicial	Dominio 3	5.849	98,32 %	0,329
Dominio 1	Inicial	Dominio 2	16.530	95,82 %	0,305
Dominio 2	Inicial	Dominio 1	13.272	96,97 %	0,291
Dominio 1	Detallado	Dominio 1	15.291	99,33 %	0,129
Los 3 dominios	Detallado	Los 3 dominios	35.760	98,38 %	0,131

Tabla 8 Resultados de la evaluación con aprendizaje de reglas

Podemos observar que los resultados son muy buenos, especialmente cuando hay homogeneidad entre el conjunto de entrenamiento y evaluación. La incorporación de un conjunto más detallado de símbolos ha hecho mejorar los resultados aunque no significativamente. Cuando el conjunto de evaluación es de un dominio distinto al de entrenamiento, los resultados caen significativamente. Cuando reunimos todos los corpora, los resultados caen significativamente, aunque significativamente menos que en el caso anterior. Parece que el sistema adolece de sobre-entrenamiento y sobre-adaptación al dominio y una capacidad de generalización mejorable.

Analizando los casos de etiquetado fuera de dominio, observamos que hay:

§ **Verbos no predichos** (Privilegie, conectaría, intenta, unifica, suprimió, envía,...), predichos como sustantivos en vez de como verbos por encontrarse en un contexto de signos de puntuación, o de infinitivo, o de pronombres (que podrían ser artículos definidos), etc. El sistema morfológico de los verbos en castellano puede llevar a engaño, especialmente en palabras cortas (‘van’).

§ **Nombres propios**, especialmente los compuestos (‘Gran Bretaña’), los que admiten determinante (‘Oeste’, ‘Mezzogiorno’) o cuando van en coordinación.

§ **Indefinidos** que pueden ser adjetivos o pronombres en función de un contexto de más de 2 palabras (‘algunos’).

§ **Locuciones**, no modeladas como macrounidades, y que provocan un contexto amplio (‘cada vez que’).

§ **La palabra ‘que’**, que puede ser conjunción o pronombre relativo en función de un contexto potencialmente muy lejano.

§ **Adjetivos/sustantivos**: difíciles de distinguir en casos de coordinación de contexto amplio (‘el juez irlandés y el italiano’), en nominalizaciones (‘lo contrario’, ‘la menor’), en pares epítetoAntepuesto-sustantivo (‘las distintas desventajas’, ‘de baja competitividad’), menos frecuentemente en los pares sustantivo-adjetivoEspecificadorPostpuesto (‘las políticas aplicadas’), en pares de adjetivos consecutivos (‘zonas urbanas pobres’).

§ Elementos de **categorías cerradas** (adverbios, numerales, conjunciones, pronombres...) no presentes en el corpus de entrenamiento.

3.2.3.3 Desambiguación contextual estocástica

El etiquetado *TnT* es un etiquetador estocástico configurable que nos permite emplear un diccionario de apoyo, además de la posibilidad de emplear su propio módulo léxico antes comentado (*T. Brants 1999*). Permite emplear bigramas y trigramas, con diferentes modos de suavizado, aunque los mejores resultados los hemos obtenido con la interpolación lineal de unigramas, bigramas (y, en su caso, trigramas).

Vamos a experimentar con los siguientes parámetros:

§ **Conjunto de etiquetas**: *ntt* será el conjunto de etiquetas completo y *smp* el conjunto simplificado.

§ **Preproceso de locuciones**: se han empleado dos estrategias para el modelado de las locuciones: una que contiene un modelado explícito de locuciones como agrupaciones de palabras; otra que modela las locuciones como secuencias ordinarias de palabras categorizadas con sus propias categorías (por lo cual el número de elementos de evaluación en cada estrategia es diferente, debido a la agrupación de los términos de cada locución en el primer caso).

§ **Corpus**: se harán experimentos con el corpus de evaluación del 860 y con el conjunto discapacidad (usado completamente como corpus de evaluación).

§ **Diccionario**: *tgb2* es el diccionario extraído a partir de la salida de nuestro módulo léxico completo (se toma el texto categorizado ambiguamente y se elabora un diccionario igualmente ambiguo) y *tg* es el diccionario de entrenamiento obtenido del corpus de entrenamiento del 860; es importante observar que en el caso *tgb2 no se emplea información de probabilidad estimada en el conjunto de entrenamiento.*

§ **n-gramas**: *n1* supone empleo de unigramas; *n2* supone interpolación lineal con bigramas; *n3* emplea trigramas e interpolación.

§ **anchura del haz** (*beam-width*): creciente entre 1 y 1000, pero 0 es el haz máximo).

3.2.3.3.1 Experimentos sobre la importancia del preprocesado de locuciones

El primer factor que analizaremos es el influjo del preprocesamiento de las locuciones, tanto como etiquetas completas como simplificadas.

Tabla 9 Gráfica de cobertura con el conjunto de etiquetas completo, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando unigramas

Usando un conjunto completo de etiquetas, unigramas y cualquiera de los diccionarios, los resultados son significativamente superiores cuando se emplea el preprocesamiento de locuciones respecto a cuando no se emplea. Respecto a los diccionarios, sólo para gran anchura (>100) consigue el módulo léxico probabilístico (*tg*) superar los resultados de nuestro módulo léxico (*tgb2*), aunque no significativamente.

Tabla 10 Gráfica de cobertura con el conjunto de etiquetas completo, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando unigramas

Analizando el preprocesamiento de las locuciones, pero con el conjunto simplificado de etiquetas, unigramas y con cualquiera de los diccionarios, los resultados son significativamente superiores cuando se emplea el preprocesamiento de locuciones frente a cuando no se emplea. Experimentando con los diccionarios, nunca consigue *tg* superar nuestro módulo léxico, aunque esta consistencia de resultados no da lugar a diferencias significativas.

Tabla 11 Gráfica de cobertura con el conjunto de etiquetas simplificadas, sin procesado especial de locuciones, sobre un conjunto de evaluación de

38.310 palabras, empleando unigramas

Tabla 12 Gráfica de cobertura con el conjunto de etiquetas simplificadas, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando unigramas

3.2.3.3.2 Importancia del contexto al desambiguar

Ahora experimentaremos con el empleo de bigramas y trigramas para resolver la ambigüedad.

Tabla 13 Gráfica de cobertura con el conjunto de etiquetas simplificadas, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando bigramas

Tabla 14 Gráfica de cobertura con el conjunto de etiquetas simplificadas, con procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando bigramas

Aunque mejoremos el modelo de lenguaje a bigramas, sigue siendo mejor el empleo de preprocesado de locuciones, tanto con etiquetas simplificadas como completas, para cualquier anchura del haz.

Tabla 15 Gráfica de cobertura con el conjunto de etiquetas completo, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.130 palabras, empleando bigramas

Tabla 16 Gráfica de cobertura con el conjunto de etiquetas completo, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando bigramas

A la vista de todos estos resultados proseguiremos el análisis considerando que se debe usar el preproceso de locuciones. Observando estos últimos resultados más en detalle, constatamos que usando *tg* filtramos mejor (mejor cobertura, menor ambigüedad). Ello es debido a que *tgb2* carece de adaptación a las probabilidades léxicas del entrenamiento. Sin embargo, al usar etiquetas simplificadas, la tendencia de la cobertura se invierte y es consistente para diversas anchuras del haz de búsqueda, aunque la ambigüedad de *tg* sigue siendo menor. Como los resultados con etiquetas simplificadas son significativamente mejores, proseguimos con esta opción.

Tabla 17 Gráfica de cobertura con el conjunto de etiquetas simplificado, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando trigramas

Al emplear un modelo de lenguaje superior no se obtiene una mejora consistente: los trigramas se comportan mejor cuando el filtrado que se busca es más fuerte, mientras que los bigramas son mejores significativamente cuando se amplía la anchura de la salida (siempre aplicando reducción del espacio de búsqueda en haz).

Tabla 18 Gráfica de cobertura con el conjunto de etiquetas simplificado, con procesado especial de locuciones, sobre un conjunto de evaluación del dominio de discapacidad de 22.518 palabras, empleando bigramas

Tabla 19 Gráfica de cobertura con el conjunto de etiquetas simplificado, con procesado especial de locuciones, sobre un conjunto de evaluación del dominio de discapacidad de 22.518 palabras, empleando trigramas

3.2.3.3.3 Influjo del cambio de dominio

Si procedemos a cambiar de dominio pero mantenemos el entrenamiento anterior, los resultados favorecen la opción de un módulo léxico general sin probabilidades adaptadas.

Las tendencias se mantienen al usar trigramas en vez de bigramas.

3.2.4 Conclusiones sobre etiquetado automático

En el primer módulo (preprocesamiento), donde no disponíamos de base de datos de aprendizaje, hemos obtenido resultados excelentes por medio de reglas manuales, técnica muy eficaz aunque inicialmente costosa, que puede servir para crear una base de datos para entrenar técnicas basadas en aprendizaje. Destacamos que las reglas de la división en sílabas se han mostrado muy precisas en la detección de palabras al escritas y expresiones extranjeras.

En cuanto al módulo de información léxica, que debe incorporar la máxima y más precisa cantidad de información gramatical, hemos demostrado que es adecuada una combinación formada por unos diccionarios amplios, un completo subsistema de conjugación verbal y unas precisas reglas de terminaciones, superando en cobertura los resultados de un módulo estocástico.

En la etapa de desambiguación contextual hemos ensayado tres técnicas (reglas manuales, reglas inferidas automáticamente y aprendizaje estocástico), entre las cuales sobresalieron las dos últimas, la técnica estocástica y el aprendizaje automático de reglas.

En la técnica estocástica, al comparar los resultados de un módulo léxico generalista y no adaptado a un dominio (sin probabilidades) con el módulo léxico adaptado al dominio de entrenamiento, no se encontraron diferencias significativas hasta que se experimentó con un cambio de dominio de evaluación (sin cambiar el dominio de entrenamiento), que mostró la superior capacidad de generalización del módulo no probabilístico.

Los mejores resultados de desambiguación se han conseguido con interpolación lineal de n-gramas, sin que haya diferencias significativas en el empleo de bigramas o de trigramas como n-grama máximo, sin duda por la interpolación lineal y el tamaño mediano del corpus.

Comparando con otros autores en castellano, los resultados de esta tesis son superiores tanto en la parte léxica como en la desambiguación, aunque las condiciones de experimentación diferentes (corpus y etiquetado) no nos permiten en general extraer conclusiones significativas, sólo indicios. Si aquí se obtiene un 99,87% de cobertura con 1,6984 categorías por palabra, otras plataformas como la de ITEM alcanzan un 99,5% con un 1,64 categorías por palabra

Si aquí se obtiene un 98,75% de cobertura (sin ambigüedad), en (*E. Dermatas & G. Kokkinakis* 1995) se aplica un sistema probabilístico obtiene una tasa de error inferior al 4 % sobre las 11 etiquetas de primer nivel, e inferior al 6 % sobre el conjunto completo de etiquetas. En (*J. Zavrel & W. Daelemans* 1999) se aplican técnicas MBL sobre otro corpus (*CRATER)*, obteniendo una tasa de acierto del 97,8 %

En (*F. Pla & N. Prieto* 1998) y (*L. Márquez* 1999) alcanzaban una tasa de error del 3% que, al incluir información léxica en los bigramas, se redujo hasta un 2,58 % (*F. Pla, A. Molina & N. Prieto* 2001).

3.3 Análisis sintáctico automático y robusto

3.3.1 Análisis sintáctico

El objetivo de este capítulo es describir y evaluar una técnica de análisis sintáctico (*parsing)* robusto de cara a su futuro uso en la predicción prosódica.

Partiremos para ello de los algoritmos y programas desarrollados en (*J.M Montero* 1992) por el autor de esta Tesis, posteriormente complementados en (*D. Polanco* 2001), PFC dirigido asimismo por el autor. Concretamente emplearemos el algoritmo ascendente CYK allí descrito, en su versión no síncrona en símbolo, de complejidad cúbica con la longitud de la cadena de entrada. Aunque la robustez de las gramáticas y el procedimiento de recuperación nos permitiría emplear gramáticas regulares, hemos preferido emplear gramáticas de contexto libre porque disponíamos de herramientas para el análisis ambiguo completo de frases y no hemos tenido que recurrir a nuevas herramientas propias o ajenas. El análisis completo nos permite fácilmente la recuperación del análisis más simple, método que proponemos para combatir el sobreanálisis típico de una gramática muy robusta.

```
for i=1..N /* inicialización */
 for j=1..N
  for s=1..S
   ChartCYK[i][j][s]=0;
for i=1..N /* bucle de símbolos de entrada */
 for r=1..R1 /* bucle de reglas unarias */
  if CadenaEntrada[i]==ReglasUnitariasFNC[r][2]
   for s=1..S /* bucle de símbolos en el chart */
    if SimbolosEquivalentes[ReglasUnitariasFNC[r][1]][s]==1
     ChartCYK[i-1][i][s]=1;

for i=1..N
 for j=1..N
  for k=1..N /* bucle auxiliar de recorrido de casillas */
   for r=1..R2
    if ChartCYK[i][k][ReglasBinariasFNC[r][2]]==1 &&
     ChartCYK[k][j][ReglasBinariasFNC[r][3]==1
     for s=1..S /* bucle de símbolos de la casilla [i][j] */
  if SimbolosEquivalentes[ReglasUnitariasFNC[r][1]][s]==1
   ChartCYK[i][j][s]=1;
```

Cuadro 1 Algoritmo CYK

Como ya se ha señalado previamente, los aspectos especialmente difíciles del análisis sintáctico basado en reglas son cómo alcanzar una alta cobertura, sin exceder un tiempo razonable de proceso y manteniendo un nivel adecuado de análisis (la máxima cobertura podría conseguirse con una gramática muy simple que dice que una oración es una secuencia arbitraria de palabras y signos de puntuación, pero esta gramática resulta demasiado plana para ser de utilidad).

Abordaremos estos temas mediante un análisis en 2 niveles; un primer nivel se encargará de los aspectos más locales de la sintaxis, aquellos que tratan del ordenamiento de las palabras dentro de los sintagmas más simples, para a continuación tratar algunas de las combinaciones que permiten formar sintagmas compuestos a partir de la segmentación anterior.

```
Reconstruye (i,j,simbolo,numArbol,minArbolPrevio,maxArbolPrevio)
 {minArbolPrevioAux=MinArbolPrevio;
 maxArbolPrevioAux=MaxArbolPrevio;
 contArbol=numArbol;
 if j-i==1
   {ReconstruyeDiagonal(i,j,simbolo,numArbol,minArbolPrevioAux,   maxArbolPrevioAux)
   ContArbol=IncrementaContadorArbolesUsados();}
 else
   {for k=i+1..j-1
    for s=1..S
     for t=1..S
      for r=1..R2
       if ReglasBinarias[r][1]==simbolo &&
           ReglasBinarias[r][2]==Chart[i][k][s] &&
           ReglasBinarias[r][3]==Chart[k][j][t]
        {
        MinArbolPrevioAux= MinArbolPrevio;
        MaxArbolPrevioAux= MaxArbolPrevio;
        if numArbol!=contArbol
          {OK=CopiaArboles(NumArbol, ContArbol,
MinArbolPrevioAux,MaxArbolPrevioAux); }
         Reconstruye(i, k, Chart[i][k][s], ContArbol, MinArbolPrevioAux,
MaxArbolPrevioAux);
         Reconstruye(k, j, Chart[k][j][t], ContArbol, MinArbolPrevioAux,
MaxArbolPrevioAux);
         ContArbol= IncrementaContadorArbolesUsados();
         minArbolPrevioAux= MinArbolPrevio;
         maxArbolPrevioAux= MaxArbolPrevio;}
   MaxArbolPrevio= ContArbol-1;
   MinArbolPrevio= NumArbol;}
}
```

```
donde ReconstruyeDiagonal es similar pero busca reglas unarias en vez de binarias
```

Cuadro 2 Algoritmo de reconstrucción CYK

3.3.2 El algoritmo CYK

El algoritmo CYK en su versión estándar (Cuadros 1 y 2) requiere que las reglas estén expresadas en Forma Normal de Chomsky (FNC), esto es, reglas cuyo formato debe ser: “A->a”, o bien, “A->B C”), aunque cualquier gramática de contexto libre puede ser convertida a FNC. La existencia de reglas unitarias del tipo A->B, requieren un tratamiento aparte, estableciendo una matriz de categorías no terminales equivalentes SimbolosEquivalentes[S][S], donde cada símbolo es equivalente a sí mismo; de esta manera, al reconstruir un análisis se respetará la estructura original de la gramática tal cual la escribió el autor.

Dada una cadena de símbolos terminales CadenaEntrada[N] de longitud N, dados un conjunto de reglas unitarias ReglasUnitariasFNC[R1][2] y binarias ReglasBinariasFNC[R2][3] de tamaños R1 y R2, y un conjunto de símbolos Simbolos[S] de tamaño S, el análisis CYK se basa en una matriz tridimensional ChartCYK[N][N][S] que se rellena de acuerdo con el siguiente algoritmo (*J.M Montero 1*992):

Un adecuado ordenamiento de las reglas y de los símbolos equivalentes (o un preprocesamiento de los mismos en

forma de nuevas reglas) permite acelerar el algoritmo en sus bucles interiores.

3.3.2.1 Recuperación de todos los análisis correctos

La reconstrucción de todos los árboles de análisis posibles (o secuencias de reglas aplicables para construir la cadena de entrada), requiere recursión múltiple, dado que además de invertir el proceso de análisis (el *backtraking* típico de los procesos basados en programación dinámica como es CYK), debemos multiplicar el número de árboles cada vez que detectamos una ambigüedad (dos maneras distintas de rellenar una misma casilla de análisis).

Dada una casilla (i,j) que debe contener el símbolo no terminal 'simbolo', partiendo de que previamente hemos conseguido 'numArbol' análisis previos (aunque sólo los que se encuentran entre los números minArbolPrevio y maxArbolprevio se corresponden con el análisis que vamos a realizar) podemos construir el conjunto de subanálisis que han llevado a que se haya añadido dicho símbolo en esa casilla. El proceso es recursivo, dado que para incorporar un símbolo a una casilla ha sido necesario haber incorporado 2 símbolos en 2 casillas inferiores. Como cada símbolo ha podido ser motivado por varias reglas, cada vez que se detecta que el símbolo puede ser analizado de varias maneras es necesario crear nuevos subárboles, duplicando o copiando los hasta ahora existentes:

3.3.3 Texto categorizado y reglas léxicas

El analizador admite que el texto de entrada sea ambiguo desde el punto de vista del etiquetado, esto es, que puedan existir varias categorías asociadas a una misma palabra. Como es natural, si el texto de entrada es excesivamente ambiguo (no ha sido desambiguado o apenas lo ha sido), el despliegue combinacional de posibles árboles de análisis va a desbordar las capacidades del analizador, resultando en un proceso inviable. Por eso en el apartado de categorización gramatical hemos hecho un gran énfasis en mitigar la ambigüedad del etiquetado automático, aunque manteniendo una cierta ambigüedad residual que permita mantener una cobertura alta que no excluya las categorías que podrían llevarnos al análisis correcto, y así se ha seguido un procedimiento de categorización por filtrado.

Las categorías 860 del categorizador, aunque son de carácter general, son excesivamente ricas en algunas subcategorías (en especial por motivos semánticos, aunque esto fue paliado por medio de las simplificaciones o macrocategorías introducidas), pero poco detalladas en otras (algunas de las más usuales). En general se buscaba un etiquetado con pocas categorías formadas por una sola palabra y su familia léxica. Si pensamos en aplicaciones de reconocimiento de voz, la cadena de entrada puede contener errores de inserción, sustitución y borrado, cuyos efectos hay que minimizar. En nuestro caso el texto de entrada es altamente fiable y se puede aprovechar la riqueza de algunas palabras, aunque siempre pensando en que se desea desarrollar una gramática robusta y orientada a prosodia. A las categorías 860 se les aplicó un primer transductor léxico que permite adaptar los símbolos terminales presentes en la gramática y los símbolos presentes en el diccionario. Con ello se permite emplear una misma gramática con diversos diccionarios y diversas codificaciones, así como adaptar un mismo diccionario para su uso con gramáticas de diferente complejidad.

- **Unidades léxicas compuestas**: si se trabaja con un sistema de categorización que no junta las palabras que forman una locución pero que las detecta, o si se trabaja con uno que sí las junta, las reglas relativas a estas unidades homogeneizan su tratamiento.

- **Preposiciones**: se distinguen especialmente la preposición "de" (no marcada semánticamente y que se liga con gran fuerza al "complementable" más cercano), las contracciones (que limitan el posible uso de determinantes en el sintagma).

- **Sustantivos**: se marcan por separado los componentes de una fecha o una hora, y la palabra 'número' por construcciones especiales como "la butaca número 33".

- **Adverbios**: destacan los adverbios sintácticamente especiales como los llamados de tiempo y lugar (que admiten construcciones nominales como "de hoy" o "por aquí"), los cuantificadores y el negador.

- **Pronombres**: muchos de ellos son sintácticamente particulares, presentando distribuciones propias y comportamientos no sólo sustantivos sino también adverbiales (en palabras como "cuánto", "dónde" o la muy especial "sí") o adjetivales (en el caso de los posesivos).

§ **Adjetivos**: los llamados indefinidos y numerales se comportan frecuentemente como adjetivos muy peculiares (“yo mismo”, “ellos solos”).

§ **Determinantes**: casi todas las palabras con una cierta componente de determinación son diferentes entre sí desde un punto de vista sintáctico o de distribución: artículos determinados e indeterminados, demostrativos, posesivos, el relativo “cuyo” y el preartículo “todo”.

§ **Comparaciones**: aunque en el primer nivel no se realiza este análisis, se marcan las palabras implicadas.

§ **Signos de puntuación**: pueden unarios (coma, punto y coma, etc.) o binarios (comillas, paréntesis, etc.).

§ **Conjunciones**: se tratarán como elementos aislados en este nivel sintagmático del análisis, dejando para más adelante un tratamiento más detallado.

§ **Formas verbales**: en el 860 no se dispone de información sobre todos los verbos auxiliares (como los modales o las perífrasis), por lo que nos limitamos a “haber”, “estar” y “ser”.

3.3.4 Análisis sintagmático y reglas de corte

En el marco de un análisis sintagmático no buscamos la estructura global de la frase. Por ello el ámbito de una oración completa suele ser excesivo para el análisis (entendiendo por oración el fragmento de texto segmentado por el preprocesador), aunque se pueden emplear técnicas de compactación de grafos dirigidos acíclicos para reducir la combinatoria provocada por la ambigüedad de las distintas partes de la oración. Como el tiempo de análisis depende cúbicamente de la longitud de la frase y, como acabamos de señalar, las ambigüedades de las partes, en el peor de los casos, se multiplican entre sí para aumentar la ambigüedad global, sería muy interesante subdividir cada oración en suboraciones reduciendo el tiempo total de análisis y la ambigüedad total; si las reglas no modelan efectos de largo alcance y las reglas de corte se corresponden con barreras sintácticas formuladas en las reglas, no se producirá una pérdida de optimalidad en el análisis (hemos de recordar que CYK se basa en la técnica de programación dinámica y que las reglas gramaticales que se emplean son independientes del contexto). No se aplicarán reglas de corte si dan lugar a suboraciones de menos de 3 palabras.

Los posibles puntos de corte intraoracional son:

§ **Signos de puntuación no binarios**: las enumeraciones, inserciones y yuxtaposiciones no forman parte del nivel sintagmático.

§ **Conjunciones coordinadas**: el fenómeno de la coordinación requiere un análisis de orden superior, dado que puede haber coordinación de palabras, de sintagmas o de proposiciones.

§ **Preposiciones o locuciones prepositivas**: salvo que aparezcan tras otra preposición.

§ **Contracciones**: excepto si forman parte de una locución.

§ **Pronombres objeto**: si no aparecen antecedidos por otro pronombre objeto o el negador verbal.

§ **Verbos**: cuando no van precedidos por un adverbio (especialmente el adverbio ‘no’ o un pronombre objeto).

§ **Adverbio ‘no’**: cuando precede a una forma verbal o un pronombre objeto.

Ejemplos de algunos elemento divididos para ser posteriormente juntados (y que por lo tanto no son errores) son:

§ **Fechas**:

- /el pasado 1//de abril/
- /el 15//de mayo/de 1981/
- /el día/de Navidad/
- \el 21\\de enero\de 1982\

§ **Nombres propios**:

- \Laurentino Delgado\\de\la\Riva\
- \José Luis\\deVilallonga\
- \\la Asociación Internacional\\de Médicos\por la Prevención\\de la Guerra Nuclear\
- \el Arxiu Fotografic\\de la Música\\
- \al Centre\\de Documentació Musical\
- \El Ballet Nacional\\de España\
- \en el Teatro\\del Liceo\

Errores de segmentación intra-oración evaluados fueron:

§ **Proposición entre comillas**: \:\"\\nos ha ayudado\mucho\\sin pedir\nada\\a cambio\"\

§ **Proposición entre comillas**: \:\"\nadie\\puede quitarnos\la primacía\\de la escuela\\de tradición flamenca\\"\

§ **Adjetivo alejado**: \puntos\\de sutura \defectuosos\

§ **Coordinación de palabras**: \en cuya formación\\y\expansión\

3.3.4.1 Resultados

En 484 sub-oraciones pertenecientes a 74 oraciones, se detectaron 4 errores (0,82%), aunque hay otros 2 errores no achacables al método de corte sino al preprocesado previo:

§ **locución no detectada**: \minuto\\a minuto\

§ **nombre propio mal detectado**: \por la empresa Agfa\\-\\Gevaert\

3.3.5 Reglas gramaticales sintagmáticas

El desarrollo de reglas ha sido realizado manualmente por el autor a partir de bibliografía como (*J. Alcina y J.M. Blecua* 1975), (*M.L. Herranz y J.M. Brucart* 1987) (*E. Alarcos-RAE* 1995) y a partir de un proceso iterativo de análisis, detección de errores y corrección de las reglas, realizado sobre el subcorpus 1 y el subcorpus 2 del 860, para una evaluación final con el subcorpus 4.

El desarrollo tuvo dos fases: una sintáctica y otra sintagmática. Inicialmente se construyó una gramática generalista basada en la bibliografía disponible y en los conocimientos lingüísticos del autor. En esta primera gramática se contemplaban fenómenos complejos tales como:

§ **el sintagma modificador oracional**: en posición inicial,

§ **proposiciones subordinadas**: sustantivas, adverbiales, de relativo, de relativo con ‘cuyo’, de infinitivo, gerundio o participio,

§ **las distintas modalidades oracionales**: enunciativa, interrogativa y exclamativa,

§ **la yuxtaposición y la coordinación**: copulativa, adversativa, exclamativa o distributiva,

§ **preguntas en estilo directo**,

§ **las oraciones incompletas**: esto es, sin verbo en forma personal,

§ **elementos insertables**: expresiones parentéticas, entre guiones o comillas, y expresiones adverbiales de gran movilidad,

§ **la aposición nominal**,

§ **las estructuras sintágmáticas restringidas**: con pronombres o nombres propios,

§ **los principales sintagmas simples**: preposicional, nominal, adjetivo, adverbial y verbal.

Una gramática clásica como esta contiene realmente 2 niveles de análisis oracional: un nivel sintagmático o conceptual (sobre el que sentaremos las bases de nuestra gramática de sintagmas simples) y un nivel sintáctico o relacional (en él se analizan las dependencias entre los bloques básicos). Una vez dejamos aparte la parte sintáctica, al disponer de un corpus de referencia, procedimos a perfeccionar una gramática de segmentación sintagmática. Como el corpus no se encontraba inicialmente segmentado, en el proceso perfeccionamiento de la gramática procederemos a la segmentación del corpus de referencia.

3.3.5.1 Principales segmentos (sintagmas simples)

Como segmentos de análisis se pueden destacar los siguientes:

§ **Sintagmas nominales (SN)**: incluye las fórmulas tradicionales del sintagma nominal (estructuras en torno a un sustantivo o en torno a un pronombre que actúan como núcleos) y la nominalización (estructuras en las que un determinante provoca que una estructura no nominal pueda actuar como si lo fuese)

§ **Sintagmas preposicionales (SP)**: incluyen los sintagmas nominales encabezados por una preposición, así como estructuras especiales como las que expresan cantidad, fecha, hora...

§ **Sintagmas preposicionales con la preposición 'de' o con la contracción 'del' (SPde):** esta preposición no marcada (tiene un vago significado de pertenencia) suele dar lugar a sintagmas preposiciones de baja movilidad que van inmediatamente después del elemento al que complementan (salvo en los casos de rección verbal, en los que la estructura de casos del verbo en cuestión puede permitir que se encuentren más alejados. Este hecho va a ser empleado frecuentemente en el nivel relacional de la gramática.

§ **Núcleo verbal (VERBO)**: incluye las formas simples y compuestas del verbo, así como las principales perífrasis y la inserción de elementos adverbiales dentro de la estructura del verbo.

§ **Sintagma Adjetival (SAdj)**: incluye las fórmulas tradicionales del sintagma adjetival como atributo o modificador verbal, no inserto en un sintagma nominal (estructuras en torno a un adjetivo que actúa como núcleo).

§ **Sintagma adverbial (SAdv)**: incluye las fórmulas tradicionales del sintagma adverbial como modificador verbal o adjetival, salvo cuando va inserto dentro de la estructura del verbo.

§ **Nexos**: incluye las formas simples y compuestas de las conjunciones, signos de puntuación y relativos (SConj, SPunt y Relativo)

3.3.5.2 Secuencia de segmentos (sintagmas simples)

Cada enunciado podrá estar constituido por cualquier secuencia de cualquier número de sintagmas simples. El más variado de los clásicos (verbal, adjetivo, adverbial) son el nominal y, por extensión, el preposicional. Las variantes dependen de la presencia o no de determinantes, de nominalizaciones ejercidas por estos determinantes (se puede hablar incluso de un sintagma determinante), de la existencia de cuantificadores, etc. Para el análisis se han establecido niveles dentro de las variantes, existiendo líneas paralelas de desarrollo en función de la presencia o no de determinante. Tomando como ejemplo el sintagma nominal y el sintagma preposicional mostrados en la Ilustración 1, el nivel inferior (SN1) es el nivel que recoge las estructuras más habituales de sintagmas construidos en torno al sustantivo: <Determinante> <modificadores antepuestos> <núcleo> >modificadores pospuestos>, donde los modificadores antepuestos y pospuestos pueden ser adjetivos, numerales, indefinidos, adverbios, participios... Se ha intentado recoger las estructuras específicas de los numerales y los ordinales, de los adverbios de cantidad (más y menos), de los adverbios cualitativos (tan, muy...), y de las familias de algunos adjetivos especiales (cualquiera, misma, etc.) buscando reflejar estas estructuras gramaticales, aunque se produzca cierta sobregeneración. Estas estructuras hacen que la gramática sea más detallada, pero permitirían diferenciar estos casos en las etapas prosódicas posteriores.

Por encima de él se encuentra el nivel de sintagma básico o segmento propiamente dicho (SN o SN0), en el cual se incorporan las estructuras sin determinante y las distintas nominalizaciones.

Este nivel de sintagma nominal es la base sobre la que se forman la mayoría de los sintagmas preposicionales (SP), aunque la preposición puede actuar como nominalizador por lo cual SP puede tomar estructuras nominales de niveles inferiores al SN0 . Esta sobregeneración será posteriormente tratada por el principio de mínima longitud de descripción, como se explica más adelante.

La complejidad de las estructuras gramaticales hace que se dé el caso en el que un sintagma preposicional, que toma como base un sintagma nominal más pequeño, pueda a su vez ser nominalizado por un determinante para formar un sintagma nominal mayor (SN_SP) sin el concurso de un núcleo nominal, estando este sintagma preposicional nominalizado en el nivel inferior del sintagma nominal que lo incluye.

3.3.5.3 Filtros de concordancia

A fin de limitar el sobreanálisis a que da lugar una gramática tan ambigua como la usada, y aprovechando la disponibilidad de información sobre el género y el número de las unidades léxicas, se aplica esta última información a modo de filtro que desestime análisis no congruentes con la concordancia que dentro de un sintagma simple se debe dar en castellano.

Procediendo según el modelo de filtro no es necesario modificar el analizador para realizar esta comprobación durante el análisis (aunque de esta otra manera ganaríamos algo de rapidez); tampoco se necesita multiplicar el número de categorías y sintagmas, automática pero artificialmente, para que cada categoría y cada sintagmas identifiquen su género y su número, dando lugar a una consecuente multiplicación del número de reglas.

Ilustración 1: Esquema de los principales sintagmas analizados y sus relaciones

El encadenamiento de concordancias comenzará con la primera palabra de aquellos sintagmas que puede estar dotados de género y número, los que pivotan en torno a determinantes, sustantivo y adjetivos; a partir de esta primera palabra las demás deben ser coherentes con la información impuesta por las palabras precedentes. Si alguna de las palabras carecen del rasgo de género o número o ambos, será compatible con cualquier combinación previa y se limitará a transmitirla; si esta palabra neutra es la primera, como en el caso de las preposiciones que encabezan los sintagmas preposicionales, simplemente no impone condición a los elementos restantes, no siendo necesario considerar este caso como especial.

Existen algunas categorías que dentro de los sintagmas nominales complejos rompen la cadena de concordancias; estos son las preposiciones, las contracciones y las locuciones prepositivas.

3.3.5.4 Principio de mínima longitud de la descripción

Aun habiendo aplicado restricciones de género y número sobre unos análisis de sintagmas muy básicos, es habitual que varias hipótesis sobrevivan al filtrado y sea necesario elegir alguna de ellas como la correcta. Si se dispone de probabilidades asociadas a las lexías y a las reglas de producción es posible adaptar el algoritmo, aunque nuevamente resultaría más simple aplicarlas en post-proceso, pero. la estimación de dichas probabilidades requiere el uso de un gran corpus etiquetado con esa misma gramática.

En numerosas ocasiones se ha propuesto una característica simple que podría explicar la mayoría de los análisis correctos: la sencillez, tomada esta como número mínimo de elementos en la descripción que realiza un análisis de la cadena de entrada (lo que se conoce como Minimun Description Length Principle). El concepto de longitud de la descripción no es único, pudiéndose aplicar:

- al número de sintagmas simples presentes en un determinado texto,
- al número de reglas aplicadas para la producción del texto,
- al número de niveles de la descripción en árbol,
- al número de símbolos no terminales presentes en el árbol.

De todas estas posibilidades sólo la primera goza de cierta independencia de la gramática y su escritura (siempre que supongamos como conocido el conjunto de sintagmas simples que buscamos) y es la que se ha aplicado en el

presente análisis. Dejaré aparte las consideraciones sobre la naturaleza del lenguaje como capacidad de comunicación que podrían subyacer al enunciado de este principio (para explicar la gran capacidad para aprender una lengua de los humanos Chomsky suele recurrir al posible carácter innato de algunos de sus componentes, una de las cuales sería este principio); intuitivamente podemos justificarlo con algunos ejemplos aunque, obviamente, no podemos demostrarlo. Son muchos los elementos del lenguaje que pueden funcionar como básicos (únicos) o como adjuntos o complementos dentro de otro elementos básicos mayores:

§ **Adjetivos**: pueden ser adjuntos nominales o núcleos de sintagmas determinantes (nominalización), pero también pueden ser complementos predicativos o atributos por sí solos.

§ **Sustantivos**: es infrecuente, pero posible, que sean un sintagma nominal sin adjuntos ni determinantes.

§ **Pronombres**: aunque su capacidad combinatoria es muy simple y suelen formar ellos mismo su propio sintagma, sí que admiten la compañía de algunos adjetivos como las familias léxicas "mismo" y "solo".

§ **Verbos**: en castellano, actúan como auxiliares, pero pueden ser también ellos mismos verbos principales; también los participios pueden en general actuar como elementos de la estructura verbal, o por el contrario, funcionar como adjetivos.

Desde un punto de vista estocástico, podemos decir que, en los casos de ambigüedad, es más probable la opción más simple que la que da lugar a un mayor número de elementos básicos, o que la probabilidad de que un elemento constituya por sí solo un sintagma es muy reducida y sólo se debe recurrir a ella cuando globalmente no quede otra alternativa. También podemos considerar que en la mayoría de los dominios son menos frecuentes las estructuras retóricas que, siendo admisibles, incumplen el principio de la descripción mínima.

Es posible detectar un problema al aplicar este principio: los elementos de frontera que pueden ser asignados a un segmento o otro adyacente, porque son compatibles y concuerdan con ambos; que no producen una variación en el número total de segmentos de análisis y que, por tanto, son indistinguibles a la luz del principio.

3.3.5.5 Evaluación

En una primera validación en 967 segmentos, se detectaron 4 errores debidos a las reglas de corte (0,41%), 5 errores de cobertura (0,51%), 10 errores debidos a mala categorización (1,03%) y 9 errores de análisis (0,93%).

En otra evaluación (corpus 4 del 860), sobre un total de 5.703 segmentos, 20 errores debidos a reglas de corte (0,35%), 31 segmentos no fueron identificados por falta de cobertura (0,55%), 64 presentaban errores de categorización (1,10%) y 85 presentaban errores de análisis (1,49%):

§ **inserción por error de rasgo**:

- \es\ \V2901H.0..\ VERBO
- \un\ \D01##S.M##\ **SN**
- \carcinoma\pulmonar\ \N00##**S.F**##\A11..S.N##\ SN

§ **borrado por error de categorización:**

- \así\como\dolores\ \B03###6###**B03**###6###\N00##P.M##\ **SP**

§ **inserción por error de categorización:**

- \el\cardiólogo\ \D00##S.M##\N00##S.M##\ SN
- \español\ **N00**##S.M##\ **SN**

§ **inserción por error de rasgos:**

- \del\hospital\ \P03##N.0##\N00##S.M##\ SPde
- \Humana\ \N07##**N.F**##\ SN

§ **inserción por estructura compleja:**

- \por\un\producto\qúimico\comercial\aplicado\ \P00##N.0##\D01##S.M##\N00##S.M## \A11..S.M##\A11..S.N##\V0846S.M..\ SP
- \inadecuadamente\ \B03..N.0##\ **SAdv**

§ **inserción doble por relación con elemento insertado:**

- \de\aceite\ \P00##N.0##\N00##S.M##\ SPde
- \de\colza\ \P00##N.0##\N00##S.F##\ SPde
- \desnaturalizado\vendido\ \V0846S.M..\V0846S.M..\ **SAdj**
- \ilegalmente\ \B03..N.0##\ **SAdv**

§ **borrado por nombre propio:**

- \de\policía\Laurentino\Delgado\ \P00##N.0##\N00##S.N##\N07##..M##\N06##N.N##\ SPde

§ **inserción por palabra especial:**

- \casi\ \B08..N.0##\ **SAdj**
- \por\completo\ \P00##N.0##\A11..S.M##\ SP

Errores por incapacidad de análisis:

§ **error de categoría**:

- \,\pero\,\a\la\par\ \M07####### \C19##N.0## \M07####### \P00##N80## \D00##S8F## \N00##S8M##\
- \sobre\el\planeta\ \P00#######\D00##S.M##\N00##S.F##\
- \,\la\holandesa\ \M07#######\R02##H.F##\A11..S.F##\
- \:\la\de\ballet\clásico\ \M11####### \R02##H.F## \P00##N.0## \N00##S.M## \A11..S.M## \
- \y\la\de\danza\española\ \C02##N.0## \R02##H.F## \P00##N.0## \N00##S.F## \A11..S.F##\
- \sobre\ \P00#######\

§ **errores de las reglas de concordancia de rasgos**:

- \de\crear\una\escuela\propia\ \P00##N.0## \V0800**N.0**.. \D01##**S.F**## \N00##S.F## \A06##S.F##\

3.3.5.6 Recategorización

Como resultado del análisis sintagmático y de la aplicación del principio de mínima descripción es posible categorizar de nuevo el texto ambiguo de entrada (realmente se trata de selección o desambiguación más que de recategorización). Desafortunadamente la información aplicada en el nivel sintagmático es de la misma naturaleza (relaciones locales) que la empleada al categorizador y no se produce mejora respecto al mejor de los candidatos obtenido probabilísticamente.

3.3.6 Reglas gramaticales de segundo nivel (sintácticas)

La elaboración de sintagmas más complejos descansa sobre la creación de estructuras que agrupen sintagmas más simples en sintagmas compuestos. Con esta agrupación se pretenderá detectar algunas situaciones que pueden resultar prosódicamente relevantes. Basándonos en información morfosintáctica, esto es, en información sobre colocación relativa de los sintagmas simples y en la clasificación no estructural de las palabras, no se puede

abordar la estructura global de una oración (para lo cual es fundamental conocer la información sobre rección verbal). Los elementos estructurales a los que se puede uno aproximar de esta manera son:

- la coordinación, tanto intra-sintagma o intersintagma, pero no en el nivel de oración o cláusula,
- algún sintagma nominal más complejo como puede ser el de las fechas,
- la cuantificación en el nivel de sintagma,
- algunas estructuras comparativas básicas,
- el encadenamiento de los muy habituales complementos preposicionales basados en la preposición no marcada 'de' y su variante 'del'.

En estos elementos van a desempeñar un importante papel algunas palabras que podemos calificar como especiales porque requerirían un tratamiento de categoría única ('tanto', 'como', 'más', 'menos').

La entrada para la gramática de este nivel es la salida de la gramática sintagmática, siendo los sintagmas simples sus símbolos terminales y los sintagmas compuestos sus símbolos no terminales. Este nivel no puede introducir errores de cobertura porque en el peor de los casos se conserva el análisis del nivel anterior.

3.3.6.1 Evaluación

Sobre un conjunto de evaluación con 55 frases de longitud media igual a 31 palabras, el análisis dio lugar a unas 4 palabras por segmentos y 7 segmentos por frase, con un número máximo de 15 palabras en un segmento y se obtuvo una cobertura del 88,34% (470 /532) con 27 errores debidos a errores de categorización y segmentado previo (5,07%). La precisión fue del 87,52 % (470/537). Los errores se deben principalmente a estructuras de coordinación no contempladas, lo cual sugiere que es necesario un mayor trabajo futuro en el desarrollo de reglas, o su aprendizaje automático.

3.3.7 Conclusiones sobre análisis sintáctico

Aunque el empleo de un conjunto diferente de segmentos y un corpus también distinto dificulta la comparación, los resultados obtenidos en análisis sintagmático y en análisis sintáctico son similares en porcentaje a los mejores resultados en castellano, destacando su robustez y cobertura. Así la tasa aquí obtenida (un 96,5% de segmentos sobre un corpus de frases complejas) es superior al sistema APOLN (*A. Molina et al* 1999) al evaluarlo sobre parte de LEXESP (un corpus que es también sintácticamente complejo), aunque hay que tener en cuenta que el desarrollo de reglas se produjo sobre un corpus diferente, lo cual supone unas condiciones distintas a las de esta Tesis. Otros autores (*H. Jiménez & G. Morales* 2001) en castellano se limitan a la detección de sintagmas nominales, obteniendo resultados ligeramente superiores (en torno al 98 %) sobre una tarea más simple.

Las estrategias de segmentación de oraciones largas y la aplicación de una variante del principio de mínima longitud descriptiva han resultado ser muy exitosas de acuerdo con la evaluación realizada.

El tamaño del corpus empleado y el hecho de no hallarse inicialmente etiquetado sintácticamente, han hecho que sea adecuado para la elaboración manual de reglas de experto, siendo más bien pequeño para el aprendizaje supervisado o no supervisado de reglas, objetivo en el futuro planteable con el corpus etiquetado que se ha obtenido.

Capítulo 4 Modelado de la F0 para síntesis en dominio restringido

La calidad actual de las voces sintéticas, a pesar de los importantes avances experimentados en los últimos años, no es suficiente para aplicaciones de interacción vocal de calidad (*IVR*), en las cuales se suele preferir el empleo de cuidadas grabaciones de voz natural de locutores profesionales (preferentemente mujeres). Sin embargo, la síntesis resulta casi siempre imprescindible si el número de posibles elocuciones (o fragmentos de elocuciones) es considerablemente elevado. Un ejemplo de esto podría ser una aplicación que contuviese mensajes como los siguientes:

- De acuerdo señor **<apellido>**, ¿qué operación desea?
- El tren <tipo de tren> <población de origen> - <población de destino> sale a las <hora> y llega a las <hora>.

Se trata de oraciones compuestas por un patrón fijo y un cierto número de campos variables (señalados en el texto entre los signos < y >).

Si el número de elementos que pueden aparecer en un campo variable es reducido (tipo de tren, hora, etc.), la mejor solución es grabarlos e insertarlos dentro de la grabación patrón en el momento de la síntesis del mensaje. Si la grabación se realiza en el contexto adecuado (lo más cercano que se pueda al contexto de inserción), la calidad alcanzable puede ser indistinguible de una grabación completa original.

Si, por el contrario, la cardinalidad de los campos variables es elevada (apellidos, poblaciones, etc.), el número de grabaciones puede resultar económicamente inaceptable, incluso en el caso de que sólo se grabe un ejemplo de cada apellido o de cada población. En estas circunstancias de dominio restringido, la conversión texto a voz puede permitir gran flexibilidad al desarrollador de aplicaciones con un coste más aceptable. La calidad alcanzable en estas condiciones restringidas puede ser alta, haciendo posible la convivencia entre voz natural y sintética, especialmente si disponemos de la posibilidad de desarrollar una síntesis basada en la voz de la locutora o el locutor de las grabaciones patrón.

Las fases del proceso de desarrollo son:

- **Diseño de la base de datos de dominio restringido**: criterios y estructura.
- **Grabación y etiquetado prosódico-segmental** semiautomático de la base de datos.
- **Parametrización y análisis de parámetros y patrones** prosódicos.
- **Modelado de F0** mediante perceptrones multicapa.

4.1 Diseño de la base de datos de dominio restringido

Las basen de datos constituyen los cimientos sobre los que van a crecer las etapas posteriores de análisis y modelado que nos deben conducir a la definición de una nueva voz (en su parte prosódica, fundamentalmente). La calidad de estos datos de partida definirá y limitará la calidad que finalmente podamos alcanzar en el modelado prosódico. Es necesario cubrir todos los posibles fenómenos que se puedan dar en la realidad y que se pueden dar en el momento de sintetizar prosodia, objetivo final del proceso.

Una base de datos de voz como la que describiremos no es útil sólo para el modelado prosódico; por sus características de diseño, puede ser el punto de partida para un modelado segmental basado en selección de unidades (síntesis guiada por datos o *data-driven*).

La definición de una base de datos para dominio restringido viene condicionada por el dominio de aplicación, en este caso la empresa *Natural Vox*, especializada en aplicaciones telefónicas de calidad, en castellano, en entorno fundamentalmente bancario y de información de tráfico. Esta empresa necesitaba una voz femenina de alta calidad y seleccionaron al GTH para este trabajo de investigación y desarrollo porque consideraban a BORIS la mejor síntesis en castellano del mercado. Dado que el proyecto entroncaba con algunos de los objetivos del GTH en I+D (aplicaciones telefónicas, síntesis de voz), el proyecto era sumamente interesante en sí, y además permitía la definición y captura de una base de datos etiquetada que podía resultar útil para investigar en conversión de voces y síntesis por selección de unidades, temas ambos punteros en la investigación actual en Tecnologías del Habla (*J. Gutiérrez-Arriola* 2001). Aunque en esta Tesis se describe principalmente el modelado prosódico asociado al

proyecto, este concluyó con el desarrollo también de la parte segmental para la cual se empleó síntesis por concatenación en el dominio del tiempo (*J.M Pardo et al* 1995).

Tres son las respuestas a las que debemos responder en el diseño de la base de datos:

1) **¿qué tipo de textos?**
2) **¿cuántas frases o palabras grabar?**
3) **¿cuáles?**

Los factores limitantes que condicionan nuestro diseño son:

- **El dominio** de las aplicaciones de *Natural Vox*.
- **El presupuesto** limitado.
- **La experiencia previa**.
- El **objetivo** de conversión texto a voz **de alta calidad.**

En el apéndice "A.2.1 Frases patrón iniciales de la base de datos de dominio restringido" se muestran las frases patrón inicialmente definidas por la empresa. Se trata de frases y vocabulario típicos de una aplicación telefónica como las que constituyen nuestro objetivo. Son 22 frases portadoras, enunciativas o interrogativas, que contienen 30 campos variables que podríamos clasificar de la siguiente manera:

- **Nombres propios**: poblaciones, puertos de montaña, nombres y apellidos.
- **Expresiones de naturaleza bancaria**: movimientos, tipos de cuentas, divisas, etc.
- **Nombres de entidades financieras**: bancos y cajas.
- **Horas**.
- **Números de teléfono**.
- **Combinaciones de letras y números**.

Una vez analizadas las frases patrón, se realizaron los siguientes agrupamientos en función de similitudes prosódico-fonéticas:

- **Frases 1, 2, 3, 4, 5, 22**: los **nombres propios** de poblaciones y puertos de montaña **en oraciones enunciativas** y posición entre-pausas pueden ser agrupados dando lugar al campo variable <nombre propio enunciativa>.
- **Frases 6, 8**: los **nombres propios de persona** (incluyendo la combinación: nombre de pila + apellidos) **en oraciones enunciativas** y posición entre pausas, pueden ser agrupados junto a los anteriores, ampliando el ámbito de <nombre propio enunciativa>
- **Frase 20**: los **nombres propios de persona en oraciones enunciativas** y posición final de frase y entre-pausas, se pueden agrupar dentro del campo <nombre propio enunciativa>, haciéndose observar la necesidad de que las grabaciones se realicen con prosodia neutralizada (sin énfasis especial debido a estar en posición final de frase).
- **Frase 7**: los **nombres propios de persona en oraciones interrogativas** y en posición final de frase entre-pausas, dan lugar al campo variable <nombre propio interrogativa>. Dado que este campo es generalizable, consideramos necesaria la ampliación del corpus de prosodia con la inclusión de nombres propios completos y nombres de poblaciones y puertos, en oraciones interrogativas y posición final.
- **Frases 9, 10**: los **movimientos bancarios** responden a una sintaxis no habitual en el lenguaje escrito o hablado (se trata de una variedad dialectal de tipo jerga); presentan unas estructuras sintácticas simplificadas, basadas en secuencias de sustantivos, importante omisión de elementos de enlace (preposiciones) y escasez de adjetivos (generalmente forman compuestos junto a algún sustantivo). Por esta razón pasaron a formar parte del campo variable de los compuestos: <sintaxis

simple enunciativa>. No se debe olvidar que este campo de compuestos puede incluir nombres propios.

§ **Frase 11**: las **palabras clave de un servicio telefónico bancario** presentan características similares a los movimientos bancarios y podrían ser integrados en el campo <sintaxis simple enunciativa>. De nuevo recalcamos la necesidad de que las grabaciones se realicen con prosodia neutralizada (sin énfasis especial debido a ser final de frase).

§ **Frases 14, 17**: los **tipos de cuentas y nombres de entidades** implican igualmente la lectura de nombres compuestos de acuerdo con una sintaxis restringida (no incluimos las entidades dentro del campo de nombres propios debido a la posibilidad de tener que sintetizar entidades tales como *Sindicato de Banqueros de Barcelona*, que es más un compuesto con nombre propio que un nombre propio). Por ello formarán parte del campo <sintaxis simple enunciativa>.

§ **Frases 16, 19, 21**: los campos variables contenidos en todas estas frases tienen en común que pertenecen a **frases patrón interrogativas** y se encuentran en posición final de frase. Por ello los agruparemos dentro del campo <sintaxis simple interrogativa>. A fin de completar este campo de cara a futuras aplicaciones, planteamos su ampliación con la inclusión de compuestos (movimientos bancarios, etc. sin nombres propios), en oraciones interrogativas y posición final.

§ **Frase 18**: esta **oración interrogativa** presenta su campo variable no en la posición final sino en una intermedia, y es posible que su prosodia esté más cerca de la enunciativa que de la interrogativa.

§ **Frase 4**: además de lo antes reseñado, requiere síntesis de **horas** en posición final o no final, aunque creemos que compensa la utilización de técnicas de concatenación de mensajes pregrabados sin modificación prosódica. Se trata de un caso muy genérico (sirve para gran cantidad de aplicaciones) y muy acotado (el concepto de horario no está sujeto a los vaivenes del tiempo ni de las aplicaciones).

§ **Frase 12**: el **estado de un cheque** podría ser incorporado al tipo <sintaxis simple enunciativa>, aunque creemos que un campo con tan poca variedad de vocabulario puede ser tratado mediante mensajes pregrabados.

§ **Frase 13**: el tratamiento de **números de teléfono** podría ser igualmente un caso de reproducción de mensajes pregrabados ya que resultan de utilidad general para numerosas aplicaciones. Por ello desaconsejamos la generación de una prosodia adaptada a este campo variable acotado y genérico.

§ **Frase 15**: sintetizar **una letra y un número** es una tarea que puede ser fácilmente resuelta mediante mensajes pregrabados, así que desaconsejamos la obtención de una síntesis prosódica específica para este campo variable tan limitado en variedad.

Los elementos *Horas, Números de teléfono y Combinaciones de letras y números* fueron descartados en la fase inicial del diseño (frases 4, 13 y 15). Es mucho más sencillo emplear concatenación de palabras o expresiones pregrabadas (con variantes de entonación), que desarrollar un conversor texto a voz completo que alcance la calidad equivalente.

Tras esta primera depuración, las 19 frases restantes se pueden renumerar y agrupar de acuerdo con los siguientes tipos de campos (apéndice A.2.2 “Frases patrón definitivas de la base de datos de dominio restringido”):

§ **<nombre propio enunciativa>**: hay que grabar una base de datos con nombres de puertos, personas, etc., en oraciones enunciativas y en posición entre-pausas (frases 1-7, 17 y 19).

§ **<nombre propio interrogativa>**: aunque inicialmente no se trataron estas interrogativas por motivos de tiempo de desarrollo, y por no estar presente en el dominio originalmente definido, finalmente la frases 13, 16 y 18 contienen ejemplos de este tipo.

§ **<sintaxis simple enunciativa>**: hay que grabar una base de datos con sintagmas nominales (con sintaxis no telegráfica, no excesivamente concisa) de variada complejidad, en oraciones enunciativas y posición entre-pausas (frases 8-12 y 14).

§ **<sintaxis simple interrogativa>**: hay que grabar una base de datos con sintagmas nominales de

variada complejidad, en oraciones interrogativas. Hubo que ampliar la base de datos con nombres de pueblos (para las frases 13, 15, 16 y 18), aunque la frase 15 no posee un campo variable en posición final y posiblemente debería ser tratada casi como una enunciativa.

Adicionalmente y para facilitar el modelado, se estableció que las grabaciones de los campos variables dentro de las frases portadoras, se debían hacer entre-pausas (focalizando o destacando los campos variables dado que se trata de información clave), aunque debían ser leídas con naturalidad.

El apéndice "A.2.2 Frases patrón definitivas de la base de datos de dominio restringido"muestra el conjunto final de frases portadoras, donde se han recortado algunas partes de la frase portadora para hacer más rápido el proceso de grabación, alterando muy levemente la prosodia de los campos variables que se desea modelar.

4.1.1 Criterios de selección del contenido de los campos variables

Como resultado del proyecto Onomástica (*The Onomastica Consortium* 1995) se disponía de una base de datos textual con 30232 nombres de poblaciones (pueblos o ciudades), 8736 nombres de pila simples y 49431 apellidos simples. La empresa proporcionó 255 nombres de entidades financieras, 246 nombres de puertos de montaña, 23 nombres de operaciones bancarias, 150 nombres de movimientos bancarios, etc. Sin embargo, por consideraciones económicas y de tiempo, se decidió grabar solamente unas 600 frases de cada tipo.

Las bases de datos previamente grabadas en el grupo eran diseñadas manualmente por un grupo de expertos lingüistas. Por ejemplo, la base de datos para modelado de la prosodia general del castellano (*J.A. Vallejo* 1998), basada también en un único locutor (en modo lectura neutra), fue diseñada en 2 partes:

1) **Tres textos publicados** (de origen oral: un discurso y dos entrevistas) que poseen una importante variedad de esquemas entonativos.

2) **Un conjunto de frases de laboratorio** de una o dos tónicas y hasta ocho sílabas, que complementan a los anteriores.

La selección de textos ha tenido en cuenta los distintos tipos de grupos fónicos (incluyendo una cierta cantidad de ejemplos en las modalidades interrogativa y exclamativa, así como construcciones sintácticas parentéticas y enumeraciones), los distintos tipos de palabras acentuadas (oxítonas, paroxítonas y proparoxítonas) y de los distintos tipos de tonemas finales (ascendentes y descendentes). La generalidad y amplitud del dominio dificultan un tratamiento matemático de selección basada en la optimización de uno o varios criterios con búsqueda exhaustiva.

Entre los factores que influyen en la síntesis de la prosodia de un texto general, y para los cuales podríamos desear cubrir todos los posibles valores intentando reproducir la distribución de probabilidad de nuestra base de datos, destacan los siguientes:

- **Un fonema y su contexto** (clase de los fonemas anterior y siguiente).
- **¿Está en sílaba acentuada?**
- **¿Pertenece a un diptongo?**
- **¿Está en sílaba abierta?**
- **¿Está en palabra función?**
- **Posición del fonema dentro de la sílaba**.
- **Número de fonemas de la sílaba.**
- **Posición del fonema dentro de la palabra.**
- **Número de fonemas de la palabra**.
- **Posición del fonema dentro del grupo fónico.**
- **Número de fonemas del grupo fónico.**
- **Tipo de grupo fónico**: sólo tendremos 2 tipos: enunciativo entre pausas e interrogativo.

§ **¿Está en posición inicial de grupo fónico?**

§ **¿Está en posición final de grupo fónico?**

§ **Tipo de acento de la palabra** : oxítono, paroxítono o proparoxítono.

§ **Distancia silábica entre acentos**.

4.1.2 Simplificación de los criterios

Seleccionando adecuadamente los ejemplos y empleando pocas grabaciones, se puede conseguir recoger gran parte de la riqueza de los casos posibles:

§ **Prosódicos**: cubrir la mayoría de los fenómenos prosódicos que se podían dar, empleando un número reducido de ejemplos:

Duraciones: variedad fonética, silábica, de longitudes de palabra, etc.

Entonación: variedad de tipos de acentuación y de distancia entre acentos.

§ **Segmentales**: tener variedad, ya que podía llegar a ser usada como fuente de unidades:

- **Nuevos difonemas**: debido a que los iniciales no fuesen adecuados (por ejemplo, si fuesen **demasiado** breves).
- **Síntesis por selección de unidades**: por problemas de calidad segmental insuficiente, podríamos tener que recurrir a ella.

Sin embargo resulta imposible cumplir simultáneamente tantas condiciones con un número muy reducido de grabaciones. Dado que no se puede reproducir una distribución de probabilidad general (y equilibrar tantas variables) con tan sólo unas 600 frases, fue necesario reducir los criterios de selección a sólo seis criterios complejos basados en trabajos previos sobre modelado prosódico en castellano (*J.A. Vallejo* 1998) y (*R. Córdoba* 1999).

§ **Criterio fonético**: se debe intentar conseguir una distribución fonética (probabilidad de aparición de cada fonema) que no se aleje más de un 5% (como máximo) de la distribución original de la base de datos de nombres propios de que disponemos.

§ **Criterio silábico**: se debe reproducir la distribución de sílabas acentuadas / no acentuadas, abiertas / cerradas, con diptongo / sin diptongo, en posición final / en posición no final, con el criterio del 5% de desviación máxima.

§ **Criterio acentual**: se busca una distribución adecuada de palabras acentuadas / palabras función, oxítonas / paroxítonas / proparoxítonas.

§ **Criterios de palabras**: dado que los nombres pueden ser compuestos, hay que buscar reproducir la base de datos original en número de palabras por nombre propio y número de sílabas por palabra.

4.1.2.1 Simplificación para la base de datos de nombres propios

Se resumieron las listas completas del GTH y *Natural Vox* (puertos, pueblos y apellidos). En el caso de los apellidos se mezclaron criterios de selección probabilística (apellidos más frecuentes en las guías telefónicas españolas) y de resumen automático: 660 frases.

§ **Frases 2 y 3**: 3 campos variables y 50 puertos por campo dan lugar a 150 puertos contenidos en 100 frases.

§ **Frases 6, 7 y 17**: 3 campos variables y 360 apellidos (150 apellidos simples resumidos, 130 apellidos simples muy frecuentes y 80 apellidos compuestos con la palabra “de” y sin ella: 80 apellidos simples muy frecuentes + 80 apellidos simples resumidos), contenidos en 360 frases.

§ **Frases 1, 4, 5 y 19**: 5 campos variables y 50 pueblos por campo dan lugar a 250 pueblos contenidos en 200 frases.

4.1.2.2 Simplificación para la base de datos con sintagmas nominales en oraciones enunciativas

Se grabaron las listas completas asociadas a los campos variables, salvo en el caso de los bancos (resumen automático + bancos de nombre no castellano elegidos manualmente): 458 frases:

§ **Frases 8 y 9**: 4 campos variables y 150 movimientos bancarios (36 + 3 por 38), contenidos en 74 frases (36 + 38).

§ **Frase 10**: 1 campo variable y 23 operaciones bancarias, contenidos en 23 frases.

§ **Frase 11**: 1 campo variable y 7 estados de cheques, contenidos en 7 frases.

§ **Frase 12**: 2 campos variables, 43 tipos de cuentas y tarjetas y 172 nombres de bancos (157 nombres castellanos resumidos +15 no castellanos elegidos manualmente), contenidos en 172 frases (43+43+43+28+15).

§ **Frase 14**: 1 campo variable y 31 fondos de inversión, contenidos en 31 frases.

4.1.2.3 Simplificación para la base de datos con sintagmas nominales en oraciones interrogativas

Se grabaron las listas completas asociadas a los campos variables, ampliándolas con movimientos bancarios (manualmente escogidos de tal manera que el número de palabras contenido no sea superior a 3 en cada movimiento seleccionado, cuidando de no repetir las palabras más frecuentes en este tipo de textos), completando con apellidos y pueblos seleccionados automáticamente: 600 frases:

§ **Frase 13**: 1 campo variable, 43 tipos de cuentas y tarjetas, 34 movimientos seleccionados y 123 apellidos resumidos, contenidos en 200 frases.

§ **Frase 15**: 1 campo variable, 31 fondos de inversión, 46 movimientos seleccionados y 123 apellidos resumidos, contenidos en 200 frases.

§ **Frase 16 y 18** : 2 campos variables, 10 divisas, 20 tipos de información, 48 movimientos seleccionados y 122 apellidos resumidos, contenidos en 200 frases.

4.1.3 Algoritmo de selección

El problema de búsqueda presenta una complejidad exponencial:

§ **Pueblos**: la mejor de las combinaciones de 30.232 elementos tomados de 250 en 250.

§ **Apellidos**: la mejor de las combinaciones de 49.431 tomados de 150 en 150.

Se parte de que sabemos cuántos ejemplos tenemos y cuántos queremos seleccionar. En cada paso se busca minimizar localmente una distancia (o maximizar una medida de bondad), confiando en que ello no le llevará muy lejos del máximo global (*hipótesis optimista*). Pero, ¿qué distancia usar para seleccionarlos?. Analicemos algunas opciones:

§ **Energía**: es mejor elegir primero aquellos elementos que más acercan la distribución acumulada a la distribución deseada. El problema con este tipo de distancia es que se acerca muy rápidamente a cubrir la mayoría del objetivo, pero luego es incapaz de afinar los detalles y cubrir los sub-objetivos menores.

§ **Correlación**: se elige primero aquellos ejemplares que más se parecen (en términos de producto escalar de vectores-distribución) a la distribución remanente (la distribución objetivo una vez hemos excluido la distribución de los elementos elegidos hasta ese momento). También se tiende a cubrir antes aquellas componentes dominantes, marginando a las más infrecuentes.

§ **Correlación normalizada**: se intenta corregir el defecto anterior normalizando el producto escalar por la energía de cada vector.

§ **Distancia con penalización**: a fin de evitar que se elijan en exceso aquellas componentes cuyo

objetivo ya ha sido alcanzado, añadir ejemplos que contengan estas componentes se ve penalizado si se está en el entorno del 5 por ciento del objetivo final.

§ **Distancia a un objetivo parcial proporcional al objetivo final**: en vez de buscar directamente el objetivo final, se divide la búsqueda en varios sub-búsquedas (por ejemplo, 10 fases). De esta manera se va alcanzando la distribución deseada de una manera gradual.

El problema es tan complicado como interesante: la mochila multidimensional o *multidimensional knapsack* (*G. Brassard & P. Bratley* 1996). La mejor distancia encontrada ha sido la energía del error respecto a un objetivo parcial proporcional, con penalización.

Como podemos ver en el cuadro 3, en cada iteración (o paso) debemos calcular la distribución que será el objetivo en este paso; dicha distribución, si suponemos 10 pasos intermedios para alcanzar el objetivo final, se calculará como la distribución final deseada multiplicada por el número del paso (entre 1 y 10) dividido por 10. De esta manera moderaremos el carácter voraz del algoritmo, lo cual hará menos probable que busque consumir más rápidamente algunos de los componentes de los vectores de distribución, dejando otros sin cubrir totalmente.

Dentro de una iteración, debemos recorrer el conjunto de ejemplos disponibles y seleccionar el ejemplo óptimo; para cada ejemplo se ha de calcular cómo contribuye este ejemplo a alcanzar el objetivo final (teniendo en cuenta los ejemplos ya escogidos previamente). Si en algún componente del vector multidimensional el ejemplo seleccionado supone rebasar el objetivo final o intermedio, penalizaremos la selección de esta palabra, dificultando mucho su elección como óptima en esta iteración (aunque podría ser escogida en el paso siguiente cuando se cambie el subobjetivo intermedio).

§ **Calcular la distribución que será el objetivo en este paso (distribución óptima=**

§ **criterio o distribución global final* número del paso / 10)**

Dentro de una iteración, en cada recorrido por el conjunto de ejemplos disponibles:

§ **distribución local= distribución actual;**

§ **distribución local = distribución local-distribución de la palabra;**

§ **coordenada.valor= mínimo relativo (distribLocal, criterioGlobal);**

§ **distribución local = distribución local - distribución óptima**

§ **si algún valor del vector distribución óptima es menor que 0, debemos penalizar este ejemplo**

- **distancia= valor de castigo * coordenada.valor;**

§ **si no,**

- **distancia = -sumatorio del valor absoluto de las componentes del vector distribución local**

Cuadro 3 Algoritmo voraz de selección de ejemplos para una base de datos de voz

En cada iteración se selecciona el ejemplo que da la menor distancia y se incrementa el número de pasos hasta que se alcance el número de ejemplos.

4.1.4 Resultados

Vamos a ver la aplicación del algoritmo a la selección resumida de algunos problemas relativos a los dominios restringidos planteados.

4.1.4.1 Ejemplo de selección de 100 pueblos

4.1.4.1.1 Objetivo

El objetivo de la selección es resumir unos 32.000 nombres de poblaciones en sólo 100, buscando mantener los 6 vectores de distribución:

§ **Número de palabras por cada elemento variable (hasta 5 palabras por elemento)**: 53 elementos de una palabra, 18 de dos palabras, 19 de tres, 9 de cuatro y 1 de cinco.

§ **Número de silabas por palabra (hasta 5 sílabas por palabra):** 53 palabras de una sílaba, 48 de dos sílabas, 56 de tres sílabas, 25 de cuatro y 5 de cinco

§ **Distribución fonética (42 alófonos)**

§ **Posición del acento en las palabras (4 posibilidades: no tiene acento, aguda, llana o esdrújula)**:

- 41 palabras desacentuadas, 34 palabras agudas, 110 llanas y 2 esdrújulas o sobreesdrújulas

§ **Número de fonemas por palabra (entre uno y doce)**:

- 1 palabra de un fonema, 34 de dos fonemas, 17 de tres, 14 de cuatro, 27 de cinco, 27 de seis, 25 de siete, 18 de ocho, 13 de nueve, 7 de diez, 3 de once y1 de doce.

§ **Tipos de sílabas (32 tipos combinando: abiertas / cerradas, acentuadas / no acentuadas...).**

4.1.4.1.2 Selección realizada

§ **Número de palabras por cada elemento variable**: 54 18 19 9 0

§ **Número de sílabas por cada palabra**: 51 48 55 24 5

§ **Distribución fonética**:

§ **Posición del acento en las palabras**: 39 34 109 1

§ **Número de fonemas por palabra**: 1 34 16 13 26 27 25 18 13 7 2 1

§ **Tipos de sílabas**:

Podemos observar que, en términos generales, podemos conseguir un error relativo mínimo y una correlación elevada y el error absoluto medio es 2,8155%.

Vector de distribución de la selección	Error relativo	coeficiente de Pearson
Número de palabras por cada elemento variable	2,0000%	0,9999
Número de sílabas por cada palabra	2,1390%	0,9997
Distribución fonética	2,5316%	0,9996
Posición del acento en las palabras	2,1390%	0,9998
Número de fonemas por palabra	2,1390%	0,9995
Tipos de sílabas	4,5147%	0,9990

Tabla 20 Resultados de selección de 100 pueblos

Si mirásemos componente a componente, no todas las condiciones pueden ser cumplidas por el algoritmo. En términos relativos la diferencia entre el objetivo y lo conseguido se puede consultar en la Tabla 21.

Vector de distribución de la selección	Diferencia media	Diferencia máxima
Número de palabras por cada elemento variable	0,0%	100,0000%
Número de sílabas por cada palabra	0,6121 %	4,0000%
Distribución fonética	0,2683 %	5,2632%
Posición del acento en las palabras	0,8234 %	50,0000%
Número de fonemas por palabra	0,2464 %	33,3333%
Tipos de sílabas	0,2427 %	100,0000%

Tabla 21 Errores de selección de 100 pueblos

4.1.4.2 Ejemplo de selección de 150 pueblos

Al aumentar el número de ejemplos que seleccionar, el algoritmo se comporta mejor, como era de esperar, al ser más sencillo el problema, y el error absoluto medio es 1,657%.

Vector de distribución de la selección	Error relativo	coeficiente de Pearson
Número de palabras por cada elemento variable	2,6667%	0,9996
Número de sílabas por cada palabra	0,7117%	0,9999
Distribución fonética	1,0390%	0,9999
Posición del acento en las palabras	1,4235%	0,9999
Número de fonemas por palabra	2,8470%	0,9996
Tipos de sílabas	2,8571%	0,9994

Tabla 22 Resultados de selección de 150 pueblos

Si mirásemos componente a componente, no todas las condiciones pueden ser cumplidas por el algoritmo. En términos relativos la diferencia entre el objetivo y lo conseguido es:

Vector de distribución de la selección	Diferencia máxima
Número de palabras por cada elemento variable	2,5000%
Número de sílabas por cada palabra	0.0000%
Distribución fonética	2,5000%
Posición del acento en las palabras	0,6061%
Número de fonemas por palabra	50,0000%
Tipos de sílabas	100,0000%

Tabla 23 Errores de selección de 150 pueblos

4.1.4.3 Ejemplo de selección de 250 pueblos

Vector de distribución de la selección	Diferencia máxima	coeficiente de Pearson
Número de palabras por cada elemento variable	66,6667%	0,9992
Número de sílabas por cada palabra	1,5873%	0,9999
Distribución fonética	0,6623%	0,9999
Posición del acento en las palabras	25,0000%	0,9999

Número de fonemas por palabra	0,0000%	1,0000
Tipos de sílabas	7,3171%	0,9998

Tabla 24 Resultados de selección de 250 pueblos

El error absoluto medio es 0,9005%. Analizando componente a componente obtenemos los resultados de la Tabla 24.

4.1.4.4 Ejemplo de selección de 150 apellidos

El error absoluto medio es 0,5934%.

Vector de distribución de la selección	Número de elementos seleccionados	Errores
Número de palabras por cada elemento variable	150 apellidos	0,00%
Número de sílabas por cada palabra	150 palabras	1,32%
Distribución fonética	996 fonemas	0,30%
Posición del acento en las palabras	150 palabras	0,00%
Número de fonemas por palabra	150 palabras	2,65%
Tipos de sílabas	423 sílabas	0,71%

Tabla 25 Resultados de selección de 150 apellidos

4.1.4.5 Ejemplo de selección de 60 apellidos

El error absoluto medio es 2,1065%.

Vector de distribución de la selección	Números de elementos seleccionados	Errores
Número de palabras por cada elemento variable	60 apellidos	0,0%
Número de sílabas por cada palabra	60 palabras	0,0%
Distribución fonética	398 fonemas	2,76%
Posición del acento en las palabras	60 palabras	0,0%
Número de fonemas por palabra	60 palabras	3,33%
Tipos de sílabas	169 sílabas	2,37%

Tabla 26 Resultados de selección de 60 apellidos

4.1.4.6 Ejemplos de selección con baja ratio de ejemplos disponibles

Cuando tenemos poco donde escoger (baja *ratio* de selección), el error es más alto.

4.1.4.6.1 Bancos

El error absoluto medio con 150 ejemplos es 1,3738.

Vector de distribución de la selección	Número de elementos seleccionados	Errores
Número de palabras por cada elemento variable	150 bancos	7,19%
Número de sílabas por cada palabra	668 palabras	0,60%
Distribución fonética	3364 fonemas	1,19%
Posición del acento en las palabras	668 palabras	0,45%
Número de fonemas por palabra	668 palabras	2,10%
Tipos de sílabas	1399 sílabas	1,64%

Tabla 27 Resultados de selección de 60 bancos

4.1.4.6.2 Puertos

El error absoluto medio con 150 ejemplos es 1,8307.

Vector de distribución de la selección	Números de elementos seleccionados	Errores

Número de palabras por cada elemento variable	150 puertos	1,99
Número de sílabas por cada palabra	272 palabras	1,85
Distribución fonética	1377 fonemas	1,82
Posición del acento en las palabras	272 palabras	0,74
Número de fonemas por palabra	272 palabras	1,46
Tipos de sílabas	550 sílabas	2,55

Tabla 28 Resultados de selección de 150 puertos

4.1.4.7 Errores graves de selección

Del análisis de los errores detectados podemos deducir que el error desciende con el número de ejemplos, pero no su gravedad. Lo vemos en el siguiente ejemplo con Apellidos y pueblos combinados:

Número de ejemplos	Error absoluto medio	Número de errores	Número de errores del 100%
100	1,3738%	18	2
300	0,8713%	7	2
600	0,4558%	6	1
1000	0,31610%	5	1
5000	0,12886%	5	3

Tabla 29 Resultados de selección graves (entre 100 y 5000 ejemplos de pueblos y apellidos)

4.1.4.8 Algoritmo con subobjetivos intermedios

Si en vez de buscar el objetivo directamente nos planteamos pasos intermedios, los resultados mejoran considerablemente, sobre todo si la *ratio* entre el tamaño de la base de datos que resumir y el tamaño de la base de datos resumida es baja:

4.1.4.8.1 Pueblos

Número de ejemplos	Error medio con el algoritmo en 1 paso	Error medio con el algoritmo en 10 pasos
100	2,8155%	1,2760%
150	1,6572%	1,1882%
250	0,9005%	0,8255%

Tabla 30 Resultados de selección de pueblos en 1 paso y en 10 pasos (entre 100 y 250 pueblos)

4.2 Grabación y etiquetado de la base de datos

La grabación de la base de datos se llevó se cabo en una habitación tranquila aunque no aislada acústicamente. La locutora que leyó los textos era profesional y habituada a grabar para aplicaciones telefónicas como las del dominio que pretendemos modelar.

Aunque el formato de grabación fue de 16 bits y 44 Khz., para la parametrización y el análisis se convirtieron los ficheros a 32 Khz., lo cual facilita la conversión de las muestras de voz a las frecuencias estándar de trabajo de nuestro conversor (8 y 16 Khz.), sin perjudicar sensiblemente la precisión de la segmentación o el marcado.

La calidad acústica de las grabaciones es desigual; las frases 2, 3 y 4 presentan una menor relación señal ruido y una menor riqueza de altas frecuencias, así como ciertos ruidos (papeles, una puerta, etc.). Esto no afecta a la fiabilidad de nuestro análisis prosódico, pero dificulta su uso para síntesis por selección de unidades. Tras la grabación de las frases 2, 3 y 4, se corrigieron defectos en las condiciones de grabación (sin que se volviesen a grabar estas frases en las nuevas condiciones).

La labor de marcación y etiquetado fue realizada por parte de 2 becarios a media jornada durante 2 meses, con una formación inicial durante 2 semanas. Durante las primeras semanas, se hacía una revisión del marcado de los

ficheros para corregir defectos por parte de J.M. Montero y J.M. Gutiérrez-Arriola (*J.M. Montero y J. Gutiérrez Arriola* 2000). Marcaron y etiquetaron las frases 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 14, 17 y 19: (658+296 ficheros). Con posterioridad la frase 9 fue marcada por un tercer becario. Finalmente, las oraciones interrogativas fueron marcadas por una especialista en audio de la empresa (600 ficheros).

Para llevar a cabo esta labor se empleó el programa de etiquetado y marcación PCV (*J. Sánchez* 2000), (*J. Castillo* 2000) que contiene un marcador de periodos fundamentales (*pitch epochs*) como el que se describe en (*F. Giménez de los Galanes* 1995). A partir de las marcas de F0 se extraerá automáticamente la curva de F0, a la vez que permite el empleo de esta base de datos para síntesis por selección y concatenación de unidades.

Aunque es frecuente en la bibliografía emplear marcación y etiquetado automáticos (empleando HMM o DTW y la señal de un EGG) (*D. Torre* 2001), (*M.J. Makashay et al* 2000), en este proyecto hemos empleado técnicas manuales debido a que no se disponía de un EGG ni de tiempo para poner a punto un sistema de segmentación automática. Posteriormente el autor experimentó con un módulo de segmentación automática como se describe en *(M. González del Campo* 2000).

4.3 Análisis y parametrización

Una primera audición de las grabaciones reveló la existencia de un patrón bastante homogéneo para los tonemas finales de grupo fónico.

§ **Pausas forzadas pero naturales**: las pausas posteriores a los campos variables de las frases 1, 5, 6, 7, 10, 11, 15 y 17 provocaban que la entonación final del campo variable fuese descendente. Por ejemplo, "La Nacional I tiene, en la provincia de Álava, circulación ininterrumpida en <...> entre los puntos kilométricos 15 al 20", la pausa tras el campo variable resulta natural aunque para la locutora fuese obligatoria (para más detalles sobre los textos de las frases, véase el apéndice A.2.2 "Frases patrón definitivas de la base de datos de dominio restringido")

§ **Pausas forzadas y antinaturales**: las pausas posteriores a los campos variables de las frases 2, 3, 4, 8, 9, 12, 14 y 19 provocaban que la entonación final del campo variable fuese ascendente. Con ello la locutora pretendía señalar al oyente que la frase continuaría después de la pausa obligada. Por ejemplo, en la frase 3, "La Nacional-I tiene, en Madrid, los puertos de <...> y <---> cerrados", las pausas tras los campos variables no son naturales sino forzadas por nuestras condiciones de grabación.

§ **Pausas omitidas**: las pausas anteriores de las frases 6 y 7 (delante del apellido que constituía el campo variable) no fueron realizadas por la locutora y, dada la premura de tiempo, no se repitieron.

§ **Pausas no forzadas**: cuando la locutora introducía una pausa no señalada en el texto dentro de un campo variable, la entonación era, por lo general, ascendente. Así, no es infrecuente que un pequeño sintagma ("Barranco & de Mures", "Camino & Otero", "Madrid & Bolsa", "cargo por impago & o anulación", "últimos & movimientos", "banco & árabe español", etc.) contenga una breve pausa espontánea, en la que de nuevo la locutora señala la continuidad de la frase con un pico de F0 (empleado así como marcador prosódico de continuación). Aunque las pausas sean espontáneas, su lugar de aparición está sintácticamente limitada y sirve para realzar el elemento que precede a la pausa ("Barranco", "Camino", "Madrid", "impago", "últimos", "banco").

§ **Interrogativas**: las frases 13, 16 y 18 presentan la característica subida final del tono de las interrogativas (el campo variable de la frase interrogativa número 15, no presenta tonema ascendente, puesto que el campo variable no está en posición final y sus curvas de F0 se parecen más a las enunciativas).

Un análisis más detallado revela la variedad que se oculta entre la dominante homogeneidad y regularidad de un dominio restringido:

§ **Desacentuación**: determinadas palabras típicamente tónicas presentan una F0 baja como si fueran átonas (con una frecuencia fundamental baja). Esta desacentuación es más frecuente en apellidos más frecuentes (*Pérez, Ortiz, Ramos, Suárez…*). La desacentuación es más infrecuente si las posibles sílabas tónicas están alejadas (*Cuevas del Vinayo, el Humo de Rañén, Inoja de Riopisuerga, Lara de*

los Infantes…).

§ **Énfasis o foco**: determinadas palabras son pronunciadas enfáticamente (con una frecuencia fundamental muy alta). Es frecuente en apellidos muy infrecuentes (*Pels*, *Fabraque*, *señor Alonso de Robles*, *señor Alonso Carrizosa).*

§ **Acentuación por duración**: bastantes casos de desacentuación en F0, presentan duraciones típicas de vocales tónicas.

§ **Rango y nivel de F0**: a pesar de que la mayoría de las oraciones de cada frase patrón fueron grabadas de manera consecutiva, el rango y el nivel base de F0, presentan variaciones que no parecen obedecer a ninguna intención. El nivel base de F0 puede cuantificarse en 3 niveles: <150, 150 < < 160, >160 Hz. El rango de un campo variable depende del número de palabras y de la existencia de pausas espontáneas enfatizadoras.

§ **Resetting**: a veces es posible observar caídas bruscas de F0 que no coinciden con pausas: *señora | Cholla*, *señora |Cigoña.*

§ **Niveles de acentuación variables**: una estructura sintáctica como "señor de …" puede presentar diversos niveles de acentuación (enfática, normal, desacentuación): *señór | de Palamos, señór de Recort, señór de Portíya*…La entonación final ascendente es, a veces, poco enfática (*Villa Buenca*, *Las Bellostas*…). El nivel de acentuación de una vocal puede cuantificarse en 4 niveles: <170, 170 < <210, 210 < < 240, >240 Hz.

§ **Pausas no forzadas**: a veces generan entonación final ascendente (*Alquería Jorda*)

§ **Las esdrújulas**: suelen tener entonación final descendente, incluso en frases portadora que suelen provocar un patrón ascendente.

§ **Homologaciones**: a veces se introducen pausas para evitar su aparición.

§ **Las pausas espontáneas**: suelen alinearse con sintagmas (Sintagma Nominal & Sintagma Preposicional), aunque también se dan casos intra sintagma (Sustantivo & Adjetivo).

§ **Palabras átonas**: el artículo 'un' puede hacerse tónico para enfatizar. Numerosas palabras presentan tono enfático en su sílaba final sin estar en posición pre-pausa.

§ **Declinación**: su presencia depende del número de palabras y de las átonas enfatizadas.

4.4 Condiciones generales de experimentación para el modelado de F0 mediante redes neuronales artificiales

Los experimentos que posteriormente se describen forman parte del modelado prosódico de la base de datos anteriormente diseñada y grabada. Se trata de una base de datos de dominio restringido, pero de gran vocabulario. Aunque en la fase de análisis y diseño de la base de datos hemos agrupado las frases de acuerdo con criterios lingüístico-prosódicos y de aplicación, los primeros experimentos de modelado que se llevaron a cabo revelaron la naturaleza particular de algunas de las frases (como luego veremos), así como el hecho de que cada grabación con frase portadora tenía sus peculiaridades y este debía ser un parámetro de importancia (*J. Sánchez* 2000).

4.4.1 Consideraciones generales

Al igual que en (*J.A. Vallejo* 1998), emplearemos un Perceptrón Multicapa, con función sigmoide, entrenado con retropropagación, pero esta vez sin término de momento y con tan solo una capa oculta (*R. San Segundo et al* 2000). Como unidad de trabajo emplearemos la sílaba (*J.A. Vallejo* 1998), un modelado menos detallado que si trabajásemos en el nivel de fonema o inferior, pero con buenos resultados prácticos.

A la hora de evaluar, emplearemos una estrategia de *cross-validation* y *Leave-One-Out*, dividiendo la base de datos en 10 subconjuntos; 8 de ellos serán empleados para entrenar, otro será empleado para detener la red cuando comience a sobreentrenar y, finalmente, un último subconjunto será el de evaluación. Realizando 10 rotaciones de estos subconjuntos, obtenemos 10 sub-experimentos caracterizados por tener subconjuntos disjuntos de

evaluación, permitiendo con ello aumentar la fiabilidad estadística del experimento global, dado que hemos empleado todos los datos disponibles para evaluación, estrechándose así las bandas de confianza (que dependen del número de datos de evaluación).

Los experimentos se agruparán en super-experimentos, conjuntos de experimentos en los que están o no están presentes uno o varios elementos paramétricos, y que sirve para decidir si su aportación es significativa o consistente.

4.4.2 Parámetros que se ensayarán

Los principales parámetros que nos pueden ayudar a predecir la curva de F0 se pueden encontrar en la bibliografía reseñada en el Capítulo 2 sobre el estado de la cuestión (*J.A. Vallejo* 1998). Los parámetros más clásicos (o elementos de parametrización) son:

§ **Inicial (INI):** la primera sílaba tónica y las sílabas anteriores se consideran sílabas iniciales. Los experimentos en los que se emplee este elemento para definir parámetros de entrada a la red tendrán en cuenta el tamaño de la correspondiente ventana (elemento CTX), de tal manera que si usamos el elemento INI (INI=1), añadiremos a la red 2*CTX+1 parámetros de entrada.

§ **Final (FIN):** la última tónica, las sílabas posteriores a la última tónica y la inmediatamente anterior, se consideran sílabas finales. Las mismas observaciones sobre el contexto son de aplicación aquí.

§ **Acentuada (ACENT)**: la tonicidad de la sílaba y su vecindad influyen notablemente en el valor de F0. Siempre que hablemos de tonicidad nos basaremos en la información textual de que dispongamos (con la salvedad de una lista de excepciones formada por palabras función), no en criterios perceptuales o basados en la propia curva de F0 (las sílabas desacentuadas serán consideradas como acentuadas si el texto no indica lo contrario). Este elemento de parametrización ACENT se ve afectado por el elemento CTX como en los casos anteriores.

§ **Contexto (CTX)**: es el tamaño de la ventana aplicada a Inicial, Final y Acentuada. Indica el número de sílabas a izquierda y derecha de la actual que tenemos en cuenta para procesar dicha sílaba. Puede valer 0, 1, 2, 3, 4 ó 5. Así, por ejemplo, si la variable vale 2, quiere decir que realmente la ventana de observación que se está tomando es de 5 sílabas, que son la actual, 2 a su izquierda y otras 2 a su derecha. Si vale 0, evidentemente, significa que sólo se tiene en consideración la sílaba actual y no se va a evaluar el efecto contextual que sus vecinas tendrán sobre ella. Aplicando la ventana máxima de +-5 sílabas a los 3 parámetros, el número máximo de parámetros sería 33.

§ **Número de sílabas (NUM_SIL):** las frases más largas hacen mas frecuente el fenómeno del *resetting* (subida del nivel de F0 similar a la producida tras la pause inicial, pero situada en medio del grupo fónico y sin mediar una pausa), y pueden modificar la declinación general de un grupo fónico. Este elemento de parametrización puede adoptar 3 posibles valores en los super-experimentos:

- **NUM_SIL=0:** no se codificará este tipo de información, por lo que a la red no le llegará ningún parámetro relacionado con el número de sílabas
- **NUM_SIL=1:** en este caso se usará la codificación original propuesta en la Tesis de J.A. Vallejo, de tipo termómetro, donde hay 5 bits que indican si el número total de sílabas en el grupo fónico es mayor que 0, mayor que 5, mayor que 10, mayor que 15 y mayor que 20 (al ser codificación de tipo termómetro se pueden activar varios parámetros binarios simultáneamente)
- **NUM_SIL=2:** Como los valores de referencia de la codificación anterior son un poco grandes para la mayoría de ejemplos de nuestra base de datos, en este tipo de codificación se empleará el mismo número de parámetros pero con constantes diferentes (>0, >3, >7, >10 y >15).

§ **Tipo de pausa (signo de puntuación) final de grupo fónico (TERM):** los distintos signos de puntuación con los que termina cada grupo fónico determinan el tipo de tonema final del grupo y condicionan fuertemente la curva de F0. De alguna manera, indican el tipo de grupo fónico de cada

frase de acuerdo con el análisis de las frases realizado con anterioridad al diseño de la base de datos. Se consideran cuatro signos de puntuación: el punto (“.”), la coma (“,”), el punto y coma (“;”) y el guión (“-”). La codificación de esta variable es de tipo binario: cada signo de puntuación tiene un bit asignado y sólo se puede activar uno de ellos cada vez. Este elemento sólo puede adoptar los valores 0 y 4 en los super-experimentos:

- **TERM=4:** da lugar a experimentos que incluyen 4 parámetros binarios que codifican qué signo de puntuación está presente .
- **TERM=0:** estos experimentos no emplean este tipo de información.

§ **Tipo de terminación del grupo:** pueden ser cadencias o anticadencias. Como en esta base de datos el tipo de terminación está codificado según el tipo de pausa, no emplearemos este parámetro en nuestra experimentación. Los grupos fónicos terminados en coma serán mayoritariamente ascendentes (pausa espontánea de continuación), los terminados en punto serán descendentes; los terminados en punto y coma serán ascendentes (pausas forzada de continuación) y los terminados en interrogación serán igualmente ascendentes.

4.4.2.1 Nuevas codificaciones de Inicial, Acentuada y Final

Si los 3 primeros elementos adoptan el valor 7 (INI=7, FIN=7, ACENT=7), la información sobre inicial, final y acentuada se codifica conjuntamente en 7 bits (con posibilidad de aplicarles diferentes tamaños de ventana, como siempre):

- sílaba inicial y tónica,
- sílaba inicial y átona,
- sílaba no inicial, no final y tónica,
- sílaba no inicial, no final y átona,
- sílaba final, no inicial y tónica,
- sílaba final, no inicial y átona posterior a la última tónica,
- sílaba final, sílaba no inicial y átona anterior a la última tónica.

Como se puede apreciar, en la codificación original, si una sílaba es inicial y final al mismo tiempo, puede corresponder a dos casos diferentes: única tónica o anterior a la única tónica. En esta codificación, si se da este caso, dicha sílaba se codificará como inicial.

Otra posible codificación adicional de estos 3 elementos fundamentales (a la que denominaremos 6 por convención: INI=6, FIN=6, ACENT=6) contempla 6 casos correspondientes a 6 bits:

- sílaba inicial y tónica,
- sílaba inicial y átona,
- sílaba no inicial, no final y tónica,
- sílaba no inicial, no final y átona o final,
- sílaba no inicial y átona anterior a la última tónica,
- sílaba no inicial, final y tónica,
- sílaba no inicial, final y átona posterior a la última tónica.

Tras realizar varios experimentos preliminares, se comprobó que esta última codificación funciona peor. Una de las razones es que la sílaba anterior a la última tónica se ve influenciada por la última tónica y se parece más a la zona final que a la intermedia (*J.A. Vallejo* 1998). Para solucionar este problema se creó otra codificación con siete bits (la denominada 15: INI=15, FIN=15, ACENT=15):

§ sílaba inicial, no final y tónica,

- § sílaba inicial, no final y átona,
- § sílaba no inicial, no final y tónica,
- § sílaba no inicial, no final y átona,
- § sílaba final y tónica,
- § sílaba final y átona anterior a la última tónica,
- § sílaba final y átona posterior a la última tónica.

4.4.2.2 Nuevos parámetros o elementos de parametrización

Otros elementos que merece la pena ensayar son:

- § **Número de frase portadora (NUM_FRA):** aunque las grabaciones han sido realizadas buscando el menor influjo posible de la frase portadora, experimentos preliminares sugirieron incluir este parámetro
- § **Número de palabras en el grupo fónico (NUM_PAL):** se trata de un parámetro similar al del número de sílabas, que puede resultar simultánea o alternativamente útil.
- § **Posición de la palabra en el grupo fónico (POS_PAL):** nos permite distinguir entre palabras finales y no finales, por ser al final donde aparecen las cadencias y anticadencias.
- § **En palabra función (PAL_FUNC):** la pertenencia de una sílaba a una palabra función, o la vecindad de la misma, podría condicionar la curva de F0, y no sólo por la típica desacentuación. Se permite la aplicación de una ventana de +-1 sílaba.
- § **Sílaba final de palabra (FIN_PAL):** ciertos locutores tienen tendencia a realzar los finales de sintagma mediante subidas rítmicas de F0; los finales de palabra y la vecindad de palabras función se pueden combinar para predecir este hecho. Admite también una ventana de +-1 sílaba.
- § **Signo de puntuación inicial (TERM_ANT):** nos puede permitir distinguir entre grupos fónicos iniciales, y grupos fónicos tras pausa, o distinguir grupos fónicos que siguen a una cadencia o que siguen a una anticadencia. A diferencia del signo terminador de grupo fónico, aquí se presentan 5 posibilidades, las 4 antes señaladas, y el signo virtual comienzo de frase (entendiendo como frase el comienzo de la zona variable dentro de una frase portadora).

4.4.3 Elementos relacionados con la propia red

De la misma manera que experimentaremos con los elementos anteriores, también lo podemos hacer con el número de neuronas de la capa oculta y con el coeficiente de aprendizaje de la red.

Especialmente probaremos 2 tipos de normalización de la salida de la red, a fin de trasladar los valores en Hz al intervalo 0-1 con el que trabajará la red:

- § una normalización lineal basada en el valor mínimo y el rango, de acuerdo con la fórmula *ValorNormalizado= (ValorMedido-ValorMinimo)/rango*, donde *ValorMínimo* valdría 110 y *rango* valdría 150
- § una normalización lineal basada en *zscore* que emplea la formula *ValorNormalizado=.0,5 +(ValorMedido-media)*(0,9-0,1)/(1,96*desviación)*, de tal manera que el 95% de los valores observados en entrenamiento serán asignados al intervalo que va desde 0,1 hasta 0,9, si suponemos que la distribución es gaussiana. De esta manera los valores extraños y muy extremos estarán fuera del intervalo 0-1.

4.4.4 Organización de la experimentación

La experimentación que vamos a describir se organizará en diversos super-experimentos (con elementos variables que los diferencian), compuestos por experimentos (caracterizados por los parámetros de entrada a la

red), que a su vez recogen diversos sub-experimentos (de acuerdo con la estrategia de *cross-validation* y *leave-one-out*).

Dentro de un super-experimento ensayaremos todas las variaciones o combinaciones posibles de uno o varios elementos de experimentación (INI, FIN, CTX...), empleando posiblemente diversas configuraciones de la red (coeficiente de aprendizaje –COEF- y tamaño de la capa oculta –OCULTA-). Por ejemplo, un super-experimento que ensaye con diversos valores de los elementos INI (con los valores 0 y 1), FIN (con los valores 0 y 1) y CTX (con los valores 0, 1 y 2), manteniendo el resto de los elementos anulados, dará lugar a 8 experimentos:

§ **Experimento INI=1, FIN=1, CTX=2**: los parámetros de entrada a la red en este experimento serán 10, debido al enventanado de +- 2 sílaba expresado por CTX, y debido al empleo de los elementos INI y FIN con dicho valor de ventana. Los 10 parámetros (binarios todos ellos) serían:

1. si sílaba –2 respecto de la actual (anteanteanterior) es inicial[16].
2. si sílaba –2 respecto de la actual (anteanterior) es final.
3. si sílaba –1 respecto de la sílaba actual (anterior) es sílaba inicial.
4. si sílaba –1 respecto de la sílaba actual (anterior) es sílaba final.
5. si sílaba 0 (sílaba actual) es sílaba inicial.
6. si sílaba 0 (sílaba actual) es sílaba final.
7. si sílaba +1 respecto de la actual (sílaba siguiente) es inicial.
8. si sílaba +1 respecto de la actual (sílaba siguiente) es sílaba final.
9. si sílaba +2 respecto de la sílaba actual (sílaba siguiente de la siguiente) es sílaba inicial.
10. si sílaba +2 respecto de la sílaba actual (sílaba siguiente de la siguiente) es sílaba final.

§ **Experimento INI=1, FIN=1, CTX=1:** los parámetros de entrada a la red son ahora 6 (los centrales del experimentos anterior), debido al enventanado de +- 1 sílabas expresado por CTX.

§ **Experimento INI=1, FIN=1, CTX=0:** los parámetros de entrada a la red son ahora 2 (los relativos a la sílaba actual exclusivamente), debido al enventanado de 0 sílabas expresado por CTX.

§ **Experimento INI=1, FIN=0, CTX=2:** los parámetros de entrada a la red en este experimento serán 5 al incluirse sólo los parámetros referentes a la posición inicial de cada sílaba de la ventana.

§ **Experimento INI=0, FIN=1, CTX=2:** similar al anterior pero considerando si cada sílaba es final.

§ **Experimento INI=1, FIN=0, CTX=1:** 3 parámetros indicando si cada sílaba en la ventana de +-1 sílabas es inicial o no.

§ **Experimento INI=0, FIN=1, CTX=1:** 3 parámetros indicando si cada sílaba en la ventana de +-1 sílabas es final o no.

§ **Experimento INI=1, FIN=0, CTX=0:** el único parámetro empleado indica si la sílaba actual es inicial.

§ **Experimento INI=0, FIN=1, CTX=0:** el único parámetro empleado indica si la sílaba actual es final.

§ **Experimento INI=0, FIN=0, CTX=2 o INI=0, FIN=0, CTX=1 o INI=0, FIN=0, CTX=0:** se trata en los 3 casos del mismo experimento en el que no hay parámetros de entrada que tengan que ver con el carácter inicial o final de las sílabas (el tamaño de la ventana es irrelevante al no existir parámetros sobre los que aplicarla).

4.4.5 Estrategia de experimentación

Aunque tomaremos el experimento final de (*J.A. Vallejo* 1998) como referencia, realizaremos un proceso completo de experimentación con:

- los diversos elementos de parametrización comentados,
- el número de neuronas de la capa oculta (OCULTA),
- el coeficiente de aprendizaje, que puede variar entre 0 y 1 (COEF),
- el tipo de codificación de la salida,

dado que el dominio restringido puede hacer que no sean necesarios todos los parámetros allí usados, y para completar la experimentación con los elementos nuevos antes reseñados.

Existen diversas estrategias para recorrer un espacio de búsqueda tan vasto como el que nos ocupa. Si uno dispone de un cierto número de parámetros, puede comenzar la experimentación buscando el mejor de ellos (en ausencia de los demás). También se puede partir del experimento con el máximo número de parámetros e intentar ir eliminando uno a uno los parámetros menos relevantes o incluso ruidosos. Ambas son estrategias que buscan un mínimo local y que nos dan un resultado posiblemente subóptimo, pero con un reducido coste computacional y buenos resultados.

La presencia de parámetros relacionados y dependientes (por estar obtenidos por enventanado, por estar codificados según termómetro, etc.), introduce una nueva variante, dado que no resulta muy lógico probar cualquier combinación de ellos. Si disponemos de varios parámetros enventanados, debemos buscar no sólo la relevancia del parámetro, sino también cuál es el tamaño óptimo de dicha ventana. En las codificaciones de tipo termómetro hay que buscar la codificación óptima, no siendo razonable realizar pruebas con uno sólo de los parámetros que dan lugar al termómetro. Si varios parámetros ayudan a codificar un cierto espacio (por ejemplo, en nuestro caso los parámetros Inicial, Final y Acentuada ayudan a codificar la posible situación de una sílaba dentro de un grupo fónico), ya en las primeras pruebas podemos tomarlos como una unidad, aunque luego intentemos refinar la importancia individual para corroborar su complementariedad.

Dividiremos nuestro espacio de búsqueda en 3 partes:

- **parámetros generales:** son los mencionados en (*J.A. Vallejo* 1998), con la excepción del tipo de terminación, que en esta base de datos está codificado según el signo de puntuación final de cada grupo fónico.
- **parámetro específico:** al tratarse de una base de datos de dominio restringido, el parámetro "número de frase portadora" es el único de este tipo.
- **otros parámetros:** los ya mencionados (número de palabras en el grupo fónico, posición de la palabra en el grupo fónico, pertenencia a palabra función , sílaba final de palabra, signo de puntuación inicial).

Dentro de cada uno de estos subespacios estableceremos una estrategia de búsqueda diferente. En el espacio de parámetros generales partiremos de un experimento que incluye todos los parámetros, para a continuación analizar si se puede descartar alguno. En el caso del parámetro específico, al tratarse de sólo uno, la búsqueda será completa, partiendo de la base anterior. En el caso de los nuevos parámetros, partiendo de nuevo de la base anterior, la estrategia será la contraria a la de los parámetros generales: probaremos su incorporación uno a uno.

El descarte o la incorporación de un parámetro puede deberse a alguno de estos factores:

- **existen diferencias estadísticamente significativas** entre incluirlo y no incluirlo: dados los intervalos de confianza de los experimentos, podemos concluir que la inclusión de ese parámetro supone una mejora importante (es raro que la exclusión de un parámetro se traduzca en mejora significativa).
- **Existen diferencias consistentes:** es posible que la escasez de datos nos impida concluir que la inclusión (o no inclusión) de un parámetro sea o no estadísticamente significativa, pero si esta misma tendencia se repite a lo largo de una serie de experimentos, podemos optar por su inclusión (o su

descarte) aunque no tengamos significancia estadística.

§ **No existe ninguna tendencia ni diferencia significativa:** podemos descartar incluir el parámetro porque no parece estar aportando información útil al modelado, y por tanto no experimentaremos con la posibilidad de variación. Si se trata de un parámetro general, es preferible mantenerlo porque ese parámetro ha demostrado su importancia en otros dominios y aporta generalidad a nuestro sistema.

4.5 Experimentos sobre nombres propios en enunciativas

Este primer grupo de experimentación contiene las frases 1, 2, 3, 4, 5, 10, 11, 17 y 19. Se corresponde con nombres propios (en general palabras aisladas) en oraciones enunciativas. Se han excluido de este grupo las frases 6 y 7 (apellidos simples y compuestos) debido a que durante la fase de grabación no se separó sistemáticamente la palabra 'señor' (o 'señora') del apellido por medio de una pausa, obligando durante la fase de marcación y etiquetado a decidir si se procesaban o no las palabras 'señor' y 'señora'. Dado que el fenómeno no era sistemático, que la intención primera no era juntar 'señor' y el apellido, y que no se podía volver a grabar por causas económico-temporales, se decidió marcar algunas de las grabaciones de 'señor' y 'señora', y todos los apellidos. Al tratarse de unas grabaciones con características muy determinadas, no pareció conveniente mezclar estas frases con el resto de las de nombres propios en enunciativas. Esta intuición será corroborada por los experimentos que mostrarán que son más difíciles de modelar.

El número total de ejemplos de evaluación, usando *leave-one-out*, es 2099 sílabas.

4.5.1 Experimento de base con nombres propios

Al tratarse la nuestra de una base de datos de dominio restringido, donde domina el habla aislada en vez del habla continua, no podemos partir tal cual de los resultados previos de nuestro grupo. Debemos reestimar el contexto, la influencia de cada uno de los parámetros, el coeficiente de aprendizaje, el tamaño de la capa oculta y cómo codificar el número de sílabas.

Tamaño del contexto	Neuronas ocultas	Error absoluto	Error relativo
2	20	13,275	0,367
1	20	13,284	0,368
1	30	13,312	0,368
3	30	13,315	0,369
1	10	13,317	0,369
4	30	13,376	0,370
4	20	13,390	0,371
3	20	13,393	0,371
5	20	13,414	0,371
2	30	13,421	0,372
4	10	13,423	0,371
5	10	13,436	0,372
5	30	13,442	0,371
2	10	13,462	0,373
3	10	13,499	0,374
3	5	13,598	0,377
4	5	13,629	0,378
2	5	13,644	0,378
1	5	13,770	0,382
5	5	13,798	0,382

Tabla 31 Experimento de base de nombres propios, empleando zscore, los elementos sílaba inicial, sílaba final y sílaba acentuada con diversos tamaños de la ventana del contexto, con la codificación 1 del número de sílabas y empleando 4 bits para codificar el signo de puntuación final del grupo fónico.

Los resultados pueden consultarse en la Tabla 31, pudiéndose destacar que:

§ no emplear contexto (CTX=0) produce significativas diferencias respecto a emplearlo (CTX>0).

§ un número bajo de neuronas en la capa oculta (NEU=5) parece ser peor, aunque no significativamente. Dentro de este conjunto de experimentos, lo es de una manera consistente (dada una combinación del resto de los parámetros, el peor resultado se produce si el número de neuronas en la capa oculta es 5). Sin embargo, debemos observar que para un contexto pequeño, los resultados son comparables.

§ Los contextos mayores (CTX=3, 4 o 5) no introducen mejoras de una manera consistente (de hecho los mejores resultados se producen con un contexto 1 o 2).

Si comparamos los resultados basándonos en el error cuadrático medio y en el valor absoluto del error, las tendencias son las mismas, aunque no el orden (pero con un nivel de diferencias no significativo).

Tamaño del contexto	**Neuronas ocultas**	**Error cuadrático**	**Error absoluto**
1	10	19,170	13,262
2	20	19,476	13,397
4	20	19,408	13,414
3	20	19,420	13,420
1	30	19,326	13,425
4	30	19,417	13,427
1	20	19,276	13,428
3	10	19,366	13,439
5	20	19,382	13,439
2	30	19,520	13,439
3	30	19,479	13,461
2	10	19,418	13,468
1	5	19,388	13,484
5	30	19,465	13,505
4	10	19,527	13,525
5	10	19,531	13,582
2	5	19,711	13,726
4	5	19,726	13,813
5	5	19,616	13,814
3	5	19,632	13,840
0	30	24,806	16,940
0	20	24,936	17,054
0	10	25,108	17,245
0	5	25,476	17,790

Tabla 32 Experimento de base de nombres propios sin zscore.

4.5.2 Experimentos sobre la influencia de la eliminación del zscore en el experimento de base de nombres propios

Si repetimos los experimentos empleando transformación lineal de las salidas (en vez de *zscore*), los resultados no deberían variar sustancialmente, dado que la codificación de la F0 de salida no debería afectar a la capacidad de predicción de los parámetros de entrada.

Para proseguir los experimentos, tendremos en cuenta que:

§ debemos eliminar el contexto 0, significativamente peor.

§ no hay que eliminar 5 como número de neuronas, a la espera de que nuevos experimentos hagan definitivamente consistente la conveniencia de su descarte.

§ el coeficiente de aprendizaje debe ser superior (>0,3) en los casos peores (pocas neuronas, contexto nulo), por lo que de ahora en adelante nos limitaremos a los coeficientes 0,1 y 0,2.

Las tendencias señaladas se mantienen, como era previsible. Además, las diferencias entre usar o no usar *zscore* no son significativas.

4.5.3 Experimentos sobre la influencia de no codificar la información sobre sílabas iniciales, finales o acentuadas en el experimento de base de nombres propios

Como ya mencionamos al hablar de estrategia, tras estos resultados iniciales intentaremos comprobar hasta qué punto todos estos parámetros son necesarios o importantes. Comenzaremos eliminando uno a uno los elementos Inicial (INI), Final (FIN) y Acentuada (ACENT).

4.5.3.1 Omisión del elemento 'sílaba inicial'

Si prescindimos de información sobre el carácter inicial o no inicial de cada sílaba, obtenemos los resultados de la siguiente tabla:

Tamaño del contexto	Neuronas ocultas	Error absoluto
1	20	13,436
2	20	13,444
4	20	13,457
5	20	13,465
3	20	13,478
5	5	13,885
4	5	13,896
1	5	14,088
3	5	14,100
2	5	14,191

Tabla 33 Experimento de base de nombres propios, sin emplear el elemento 'sílaba inicial' (pero sí los elementos 'sílaba final' o 'sílaba acentuada').

Al eliminar el parámetro Inicial (y eliminar su contexto), los resultados no empeoran significativamente, pero se observa que el sistema es en cierta medida más independiente del tamaño del contexto. Se mantiene la tendencia a preferir más de 5 neuronas.

4.5.3.2 Omisión del elemento 'sílaba acentuada'

Tamaño del Contexto	Neuronas ocultas	Error absoluto
3	20	13,307
4	20	13,366
2	20	13,370
5	20	13,451
1	20	13,650
4	5	13,690
5	5	13,786
3	5	13,806
2	5	13,830
1	5	13,924

Tabla 34 Experimento de base de nombres propios sin emplear 'sílaba acentuada' (pero sí los elementos sílaba inicial y sílaba final).

Si prescindimos de información sobre el carácter acentuado o no acentuado de cada sílaba, los resultados obtenidos son los de la Tabla 34.

Al eliminar el parámetro 'sílaba acentuada', los resultados no empeoran significativamente, aunque ahora es clara la tendencia a necesitar un contexto mayor y se mantiene la tendencia a preferir más de 5 neuronas.

4.5.3.3 Omisión del elemento 'sílaba final'

Los resultados obtenidos al prescindir de la información sobre el carácter final o no final de cada sílaba, se muestran en la siguiente tabla:

Tamaño del contexto	Neuronas Ocultas	Error absoluto
5	20	13,682
4	20	13,721
3	20	13,902
2	20	14,204
5	5	14,389
3	5	14,619
4	5	14,713
2	5	15,373
1	20	15,516
1	5	16,119

Tabla 35 Experimento de base de nombres propios, sin emplear el elemento 'sílaba final' (pero sí los elementos sílaba inicial y sílaba acentuada).

Al eliminar el parámetro Final, los resultados empeoran significativamente en algunos casos (concretamente si el contexto igual a 1 o si el número de neuronas es igual a 5). Nuevamente es claro que se necesita un contexto mayor. Esto se explica porque los parámetros Inicial, Final y Acentuada no son independientes cuando la base de datos es fundamentalmente de palabras aisladas. Se mantiene la tendencia sobre el número de neuronas (>5).

4.5.3.4 Experimento de base de nombres propios omitiendo varios elementos (sílaba inicial, sílaba acentuada o sílaba final)

Si prosiguiésemos eliminando otro parámetro más (Tabla 36, Tabla 37, Tabla 38), observaríamos que los resultados empeoran significativamente, salvo en el caso de que mantengamos el parámetros acentuada y el contexto sea superior a 2 (hay que pensar que, aunque dominan las palabras aisladas, también hay nombres compuestos con varias tónicas).

Tamaño del contexto	Error absoluto
3	18,4965
4	18,5706
5	18,595
2	18,6178
1	19,3289

Tabla 36 Experimento de base de nombres propios, empleando el elemento final (no el de inicial ni el de acentuada), con 20 neuronas ocultas.

Tras estos análisis, optamos por proseguir con los parámetros Inicial, Acentuada y Final en la siguiente ronda de experimentos, dado que dan los mejores resultados con el menor de los contextos (lo cual se traduce en un sistema más sencillo, con menos parámetros de entrada y pesos que entrenar). Además, esta parametrización coincidirá con la general usada en (*J.A. Vallejo* 1998), salvo en el tamaño de la ventana, que será menor en nuestro caso, por ser menor la longitud de los grupos fónicos. El parámetro Final se revela como muy importante, lo cual es lógico si se tiene en cuenta que en estas frases hay mezcla de tonemas finales ascendente y descendente.

Tamaño del contexto	Error absoluto
4	17,829
5	18,098

3	17,989
2	18,407
1	19,558

Tabla 37 Experimento de base de nombres propios, empleando el elemento inicial (no el de final o el de acentuada) con 20 neuronas en la capa oculta.

Tamaño del contexto	**Error absoluto**
5	14,134
4	14,086
3	14,874
2	16,272
1	18,692

Tabla 38 Experimento de base de nombres propios, empleando el elemento acentuada (no el de final o el de inicial) con 20 neuronas en la capa oculta.

4.5.4 Experimentos sobre la influencia de eliminar el elemento 'signo de puntuación final' en el experimento de base de nombres propios

El siguiente experimento determinará la aportación del signo terminal del grupo fónico a la predicción de F0. La pérdida de calidad de modelado que se puede observar en la siguiente tabla es estadísticamente significativa (>95%), por lo cual no tiene sentido seguir probando a eliminar las terminaciones como parámetro de entrada. Esto era lógico puesto que en las frases que estamos empleando se mezclan finales ascendentes con finales descendentes, y el signo de puntuación final es el factor de discriminación.

Tamaño del contexto	**Error absoluto**
5	22,917
1	22,929
2	22,948
3	22,954
4	23,007

Tabla 39 Experimento de base de nombres propios sin emplear bits para codificar el elemento 'signo de puntuación final del grupo fónico', con 20 neuronas en la capa oculta.

4.5.5 Experimentos sobre la influencia de codificar el número de sílabas en el experimento de base de nombres propios

El siguiente parámetro que analizar es el número de sílabas. Como ya se mencionó previamente, la codificación original (denominada 1) estaba pensada para habla continua y grupos fónicos más largos, por lo cual ahora experimentaremos con su omisión, y con su sustitución por una codificación más adaptada a grupos fónicos más cortos (denominada 2).

Ninguna de las codificaciones del número de sílabas (ni la original ni la nueva) presenta una tendencia a la mejora clara o significativa, aunque entre los mejores resultados, el mejor emplea la codificación 1.

Tamaño del contexto	**Error absoluto**
2	13.346
4	13.450
1	13.468
3	13.544
5	13.587

Tabla 40 Experimento de base de nombres propios sin emplear el elemento 'número de sílabas', con 20 neuronas en la capa oculta.

Tamaño del contexto	Error absoluto
1	13.344
2	13.353
3	13.388
4	13.443
5	13.535

Tabla 41 Experimento de base de nombres propios empleando codificación 2 para el número de sílabas, con 20 neuronas en la capa oculta.

Por generalidad conservaremos la codificación original aunque no resulte especialmente útil en nuestro dominio restringido de nombres propios.

4.5.6 Segundo experimento de base de nombres propios: influencia de codificar el número de frase portadora

Este parámetro nos informa sobre cuál es la frase en la que se grabó el campo variable cuya prosodia se quiere modelar.

Tamaño del contexto	Error absoluto
1	12,278
2	12,293
5	12,348
4	12,370
3	12,408

Tabla 42 Segundo experimento de base de nombres propios: incluye además la codificación del número de frase portadora y 20 neuronas en la capa oculta.

Esta vez obtenemos mejoras significativas, confirmando que este parámetro específico de nuestra base de datos es importante a la hora de predecir F0. Como se reiteran algunos hechos anteriores (es peor emplear sólo 5 neuronas, es mejor emplear contextos menores o iguales que 2), los siguientes experimentos omitirán estos valores. Como es significativa la mejora, se incorpora el parámetro Número de frase.

4.5.7 Experimentos de nombres propios sobre otros parámetros

Analizaremos ahora el influjo de los parámetros restantes, probando inicialmente cuál de estos parámetros introduce mayor mejora; una vez seleccionado e incluido este parámetro, probamos cuál es el siguiente, y así seguimos hasta que no se produce mejora.

Tamaño del contexto	¿Es final de palabra?	Error absoluto
1	3	12,274
1	1	12,277
2	3	12,393
2	1	12,452

Tabla 43 Segundo experimento de base de nombres propios empleando el parámetro “es final de palabra” codificado sin ventana (valor 1) o con ventana +-1 (valor 3).

El empleo del parámetro “final de palabra” o incluso el aumento de su ventana (cuando vale 3 equivale a +-1 sílabas), no supone mejora significativa o consistente en este tipo de frases.

Tamaño del contexto	¿Codifica el número de palabras?	Error absoluto
1	0	12,275
2	0	12,387
1	1	12,399
2	1	12,498

Tabla 44 Segundo experimento de base de nombres propios con el elemento “número de palabras” codificado (valor 1) o no (valor 0).

El parámetro “número de palabras” empeora los resultados de una manera no significativa pero consistente (a igualdad del resto de los parámetros variables, es mejor que no emplear dicho parámetro). También es consistentemente mejor emplear contexto 1 en vez de 2.

Tamaño del contexto	¿Palabra en posición final?	Error absoluto
1	0	12,275
1	1	12,368
2	0	12,387
2	1	12,442

Tabla 45 Segundo experimento de base de nombres propios con el elemento “palabra en posición final” codificado (valor 1) o no (valor 0).

El parámetro “posición de la palabra” (en presencia del parámetro Final), no aporta nada significativo a los resultados, y en todo caso no es consistente (ni siquiera es consistente la mejoría al emplear contexto 1).

Con el parámetro “es palabra función” una vez más los resultados no son significativamente mejores, ni hay una tendencia consistente en los datos, aunque se confirma que un contexto igual a 1 es mejor.

Elegimos el parámetro “es final de palabra” como mejor parámetro que incorporar Para descubrir cuál es el siguiente mejor parámetro, se puede eliminar el contexto 2 (que es consistentemente peor), pero sólo el parámetro “Palabra Función” introduce algo de mejora, aunque es casi despreciable (12,0943 Hz), aunque se omiten las tablas de resultados porque no aportan información relevante.

Tamaño del contexto	¿es palabra función?	Error absoluto
1	0	12,275
2	1	12,344
1	1	12,352
2	0	12,387

Tabla 46 Segundo experimento de base de nombres propios con el elemento “es palabra función” codificado (valor 1) o no (valor 0).

Llevando a cabo todos los experimentos se puede comprobar que esta combinación es la mejor de todas (emplear los elementos acentuada, inicial y final con tamaño de la ventana del contexto igual a 2, con codificación 1 del número de sílabas, empleando 4 bits para codificar el signo de puntuación final del grupo fónico, con codificación del número de frase portadora y de los elementos “es final de palabra” y “es palabra función”), aunque existen muchísimas que no son significativamente peores. Entre ellas, una de las más simples, emplea una ventana +-1 sílabas para los parámetros Acentuada y Final, emplea sílabas, terminaciones y frase portadora.

4.5.8 Conclusiones sobre el modelado de nombres propios en enunciativas

Los resultados obtenidos son coherentes con los resultados de (*J.A. Vallejo* 1998), aunque en un dominio bastante cercano al habla aislada, los parámetros relacionados con la posición de la sílaba en el grupo fónico resultan algo redundantes entre sí y el número de sílabas no resulta muy importante. El parámetro más importante resulta ser el “signo de puntuación final del grupo fónico”, que permite distinguir elementos variables con cadencias y

elementos variables que generalmente presentaban anticadencias, y su incorporación es estadísticamente significativa.

El parámetro "número de frase portadora" introduce mejoras significativas, a pesar de que en las condiciones de grabación se intentó aislar el elemento variable de su frase portadora por medio de pausas obligatorias. Es posible que la simple existencia de diferentes sesiones de grabación (que en términos generales coincidieron con las frases portadoras) haya podido introducir importantes diferencias en la curva de F0 de las distintas frases.

Los nuevos parámetros ensayados apenas aportan mejoras; tan sólo "es final de palabra" y "es palabra función" mejoran algo pero no significativamente.

La estrategia de experimentación no exhaustiva (al ser comparada con la búsqueda exhaustiva del óptimo) ha mostrado su validez.

4.6 Experimentos sobre frases interrogativas

Aunque nuestra base de datos de dominio restringido contempla 4 frases interrogativas, en una de ellas (la 15) el elemento variable no está en posición final, siendo previsible que se diferencie claramente de las demás en las que esto no sucede. Estableceremos como conjunto de interrogativas puras el formado por las frases 13, 16 y 18.

El número total de ejemplos de evaluación, usando *leave-one-out*, es de 2018 sílabas.

4.6.1 Experimentos de base de interrogativas

Partiremos de considerar que el empleo de los 3 parámetros básicos (Inicial, Final y Acentuada) es necesario, dado que nuestros experimentos previos y los de (*J.A. Vallejo* 1998) nos han mostrado su utilidad. Precisamos estimar el tamaño óptimo de la ventana de parámetros, mediremos el influjo de la codificación del número de sílabas, y estudiaremos si es necesario emplear el parámetro "signo de puntuación final del grupo fónico" (ahora las grabaciones presentan final ascendente por ser interrogativas). El tamaño de la ventana es importante porque hay más grabaciones con varias tónicas que pueden hacer que el tamaño pequeño de los experimentos anteriores no sea el más adecuado.

Los resultados del experimento que nos servirá como base son los siguientes:

Tamaño del contexto	Neuronas ocultas	Error cuadrático	Error absoluto
2	10	18,259	13,156
1	20	18,322	13,197
2	30	18,320	13,211
2	20	18,379	13,242
1	10	18,396	13,258
1	30	18,436	13,287
2	5	18,481	13,314
3	10	18,330	13,333
3	30	18,422	13,354
3	5	18,367	13,358
4	5	18,410	13,390
4	10	18,518	13,462
5	10	18,527	13,489
5	20	18,547	13,507
3	20	18,658	13,508
1	5	18,635	13,515
4	30	18,584	13,524
4	20	18,600	13,536
5	5	18,660	13,539
5	30	18,658	13,602

Tabla 47 Experimento de base de interrogativas con zscore, empleando los elementos acentuada, inicial y final con tamaño de ventana del contexto igual a 1, con codificación del 1 número de sílabas, empleando 4 bits para codificar el signo de puntuación final del grupo fónico y con codificación del

número de frase portadora.

La tendencia vuelve a ser que son preferibles los contextos reducidos (menores que 3), y un número de neuronas que no resulte bajo (nos quedaremos con 20)

4.6.2 Experimentos sobre la influencia de la no codificación del número de la frase portadora en el experimento de base de interrogativas

Tamaño del contexto	Error absoluto
2	14,758
3	14,882
4	14,958
1	14,960
5	15,091

Tabla 48 Experimento de base de interrogativas sin codificación del número de frase portadora y con 20 neuronas en la capa oculta.

Partiendo de este experimento base (pero variando la presencia o no del parámetro "Numero de Frase portadora"), observamos que el predominio de los contextos reducidos es consistente, aunque no se trata de diferencias significativas. La tasa de ejemplos disponibles por cada peso que entrenar es baja (<10).

Si no empleamos el parámetro "Número de Frase portadora", las diferencias con el mejor caso sí son significativas, obligándonos a considerarlo de aquí en adelante. Es de observar que aunque al eliminar este parámetro parece que se necesita mayor contexto para conseguir peores resultados, no podemos concluir nada de esto.

4.6.3 Experimentos sobre la influencia de otros parámetros en el experimento de base de interrogativas

A partir de ahora reduciremos los experimentos a los contextos 2 y 3 que han sido los más consistentes.

Tamaño del contexto	¿Final de palabra?	Error absoluto
2	3	13,0931
2	0	13,1929
3	0	13,2751
2	1	13,2832
3	3	13,4096
3	1	13,4305

Tabla 49 Experimento de base de interrogativas codificando si es final de palabra sin contexto (1) o con contexto +-1 (3) y con 20 neuronas en la capa oculta.

Aunque el mejor resultado se obtiene empleando "Final de Palabra" con valor 3 (esto es , con una ventana de +-1 sílabas), la mejora no es significativa ni consistente (se cumple si el contexto es 2, pero con contexto 3 se invierten los resultados, siendo mejor en este caso no emplear "Final de Palabra")

Tamaño del contexto	¿es final de palabra?	Error absoluto
2	1	13,1648
3	1	13,1846
2	0	13,1929
3	0	13,2751

Tabla 50 Experimento de base de interrogativas codificando si es final de palabra sin contexto (1) o con contexto +-1 (3) y con 20 neuronas en la capa oculta.

Tamaño del Contexto	Número de palabras	Error absoluto
2	0	13,1929
3	0	13,2751
2	1	13,3164
3	1	13,4735

Tabla 51 Experimento de base de interrogativas, codificando el número de palabras y con 20 neuronas en la capa oculta.

Las mejoras motivadas por la incorporación de “Palabra Final” a los experimentos son no significativas y no consistentes (variaciones del número de neuronas de la capa oculta puede invertir la tendencia).

El parámetro “Número de palabras" es más bien negativo (el mejor resultado se da cuando no se usa), aunque de manera no significativo. Parece, por lo tanto, que no aporta nada a la capacidad de predicción, posiblemente porque otros parámetros lo hagan innecesario.

Tamaño del contexto	¿Palabra función?	Error absoluto
2	1	13,1226
2	3	13,1438
3	1	13,1858
2	0	13,1929
3	0	13,2751
3	3	13,3516

Tabla 52 Experimento de base de interrogativas, codificando o no la pertenencia a palabras función y con 20 neuronas en la capa oculta.

Como en los casos anteriores no hay ni consistencia ni significancia (los mejores y peores resultados se obtienen con la configuración 3 de “palabra función” (ventana de +-1).

Tamaño del contexto	Signo de puntuación inicial	Error absoluto
2	5	13,1639
2	0	13,1929
1	5	13,2322
3	5	13,2623
3	0	13,2751
1	0	13,3488

Tabla 53 Experimento de base de interrogativas, empleando o no 5 bits para codificar el signo de puntuación inicial del grupo fónico y con 20 neuronas en la capa oculta..

El parámetro “signo de puntuación inicial del grupo fónico” tampoco aporta mejoras, como se observa en la tabla anterior

Si se continúa la experimentación buscando el parámetro que mejor combina con “Final de Palabra”, no se logra una mejora respecto al mejor resultado obtenido hasta ahora.

4.6.4 Conclusiones sobre el modelado de interrogativas

Los resultados obtenidos son coherentes con los anteriores: el parámetro “número de frase portadora” introduce mejoras significativas.

Los nuevos parámetros ensayados apenas aportan mejoras; aunque con “es final de palabra” se mejora algo pero no significativamente.

4.7 Experimentos sobre frases enunciativas con sintagmas nominales largos

Aunque nuestra base de datos de dominio restringido contempla 4 frases enunciativas con sintagmas nominales largos, una de ellas (la 8) fue marcada con posterioridad a las demás, y experimentos preliminares mostraron que no es homogénea con las otras 3, y por lo tanto estableceremos como conjunto de enunciativas largas el formado por las frases 9, 12 y 14.

El número total de ejemplos de evaluación, usando *leave-one-out*, es de 2416 sílabas.

4.7.1 Experimentos de base de sintagmas nominales

Partiremos de considerar que el empleo de los 3 parámetros básicos (Inicial, Final y Acentuada) es necesario. Precisaremos estimar el tamaño óptimo de la ventana de parámetros, mediremos el influjo de la codificación del número de sílabas, y estudiaremos si es necesario emplear el parámetro "Signo de puntuación final". El tamaño de la ventana es importante porque hay muchas grabaciones con varias tónicas que pueden hacer que el tamaño pequeño de los experimentos anteriores no sea el más adecuado.

Tamaño del contexto	Error absoluto
4	17,0637
2	17,1600
3	17,2746
5	17,2766
1	19,1876

Tabla 54 Experimentos de base de sintagmas nominales, con zscore, empleando los elementos acentuada, inicial y final con tamaño de ventana del contexto entre 1 y 5, sin codificación del número de sílabas, empleando 4 bits para codificar el signo de puntuación final del grupo fónico, codificación del número de frase portadora, con 20 neuronas en la capa oculta.

4.7.2 Experimentos sobre la influencia de la no inclusión del elemento 'signo de puntuación final'

Si eliminamos la información relativa al signo de puntuación final del grupo fónico se obtiene:

Tamaño del contexto	Neuronas ocultas	Error absoluto
5	20	17,7286
3	20	17,7799
2	20	17,7808
4	20	17,8353
1	20	19,3665

Tabla 55 Experimentos de base de sintagmas nominales sin codificar el signo de puntuación final del grupo fónico.

Los resultados son peores pero no significativamente, por lo que procederemos a experimentar con este parámetro y el siguiente: "Número de la frase portadora".

4.7.3 Experimentos sobre la no codificación del número de la frase portadora

No hay ningún parámetro cuya exclusión suponga una aumento significativo del error, aunque contexto=1 es peor para cualquier valor de "número de la frase portadora" o de "signo de puntuación final".

Si analizamos la ausencia de alguno de estos 2, los resultados son consistentemente peores (cada uno de ellos da peores resultados que su opuesto independientemente de los valores de los demás parámetros).

Tamaño del contexto	Codificación de la frase portadora	Error absoluto
2	19	16,6952
4	19	16,7228
5	19	16,7796
3	19	16,8748
4	0	17,0637
2	0	17,1600
3	0	17,2746
5	0	17,2766
1	19	17,9549
1	0	19,1888

Tabla 56 Experimentos de base de sintagmas nominales con (19) y sin (0) codificación del número de frase portadora.

Tamaño del contexto	Codificación de la frase portadora	Error absoluto
2	19	17,0842
3	19	17,2280
5	19	17,2925
4	19	17,3481
5	0	17,7286
3	0	17,7799
2	0	17,7808
4	0	17,8353
1	19	18,1338
1	19	18,1349
1	0	19,3665

Tabla 57 Experimentos de base de sintagmas nominales sin codificar el signo de puntuación final del grupo fónico, con (19) o sin (0) codificación del número de frase portadora.

4.7.4 Experimentos sobre la no codificación del número de sílabas

En cuanto a la codificación del número de sílabas, aunque si el contexto es muy bajo (+-1), es mejor emplear alguna codificación que no emplearla, si el contexto es grande (4 o 5) es mejor no emplear codificación de "número de sílabas" (siempre de una manera no significativa). Para el resto de los valores, es mejor emplear la codificación 2.

Como consecuencia de esto, los siguientes experimentos emplearán la codificación de sílabas 2, emplearán "número de la frase portadora" y "signo de puntuación", y no emplearán contexto 1, 4 o 5 (estos últimos no parecen aportar nada con esta base de datos). En cuanto al coeficiente de aprendizaje, lo limitaremos entre 0,1 y 0,7.

En estas frases de mayor longitud, la incorporación de nuevos parámetros es buena, aunque no significativamente. Aunque sólo se muestras los datos para 20 neuronas en la capa oculta, se mantiene la tendencia a preferir el empleo de más de 5 neuronas, a pesar de que la tasa de ejemplos por peso es mayor en ese caso.

Tamaño del contexto	Codificación del número de sílabas	Error absoluto
2	2	16,6860
2	0	16,6952
4	0	16,7228
5	0	16,7796
5	2	16,7838
4	2	16,8317
3	2	16,8353
3	0	16,8748
5	1	16,9506
2	1	16,9703
3	1	17,0271
4	1	17,0551

Tabla 58 Experimentos de base de sintagmas nominales, con (1 o 2) o sin (0) codificación del número de sílabas.

4.7.5 Experimentos sobre otros parámetros

Tamaño del contexto	Palabra función	Error absoluto
2	1	16,5540
2	3	16,5823

Tabla 59 Experimentos de base de sintagmas nominales con codificación 2 del número de sílabas y con (1 o 3) codificación de la pertenencia a una palabra función.

Tamaño del Contexto	Error absoluto
2	16,5767

Tabla 60 Mejor resultado del experimento de base de sintagmas nominales con codificación 2 del número de sílabas y con codificación del número de palabras.

Tamaño del Contexto	Error absoluto
2	16,6811

Tabla 61 Mejor resultado del experimento de base de sintagmas nominales, con codificación 2 del número de sílabas, y con (1) o sin (0) emplear codificación del número de palabras.

Tamaño del contexto	¿Final de palabra?	Error absoluto
2	3	16,6001
2	1	16,6312

Tabla 62 Mejores resultados del experimento de base de sintagmas nominales, con codificación 2 del número de sílabas, , codificando si es final de palabra.

Tamaño del contexto	Signo de puntuación Inicial	Error absoluto
3	5	16,5977

Tabla 63 Mejor resultado del experimento de base de sintagmas nominales con codificación 2 del número de sílabas y con codificación del signo inicial de puntuación.

Aunque las diferencias son ínfimas, tras el parámetros “es palabra función”, los siguientes parámetros que más aportan son el “signo de puntuación con el que comienza el grupo fónico” y “es final de palabra”.

Tamaño del contexto	Neuronas ocultas	Error Absoluto
3	20	16,4168

Tabla 64 Mejor resultado del experimento de base de sintagmas nominales con codificación 2 del número de sílabas, con codificación del signo de puntuación anterior, con codificación de la posición de la sílaba en el final de una palabra y con codificación 3 de la pertenencia a una palabra función.

Llegados a este punto estos nuevos experimentos sugieren que cesemos en nuestra investigación sobre parámetros. Parece que ningún parámetros nuevo contribuye a mejorar la tasa. Sin embargo, esto es producto del método voraz subóptimo que hemos seguido para determinar cuáles son los mejores parámetros, y del propio entrenamiento de la red que no necesariamente alcanza un óptimo global.

Tamaño del contexto	¿es palabra función	¿Número de palabras?	¿Palabra final?	¿Final de palabra?	Signo de puntuación anterior	Error absoluto
2	1	1	1	3	5	16,3366
2	3	0	1	3	5	16,3372
2	0	0	1	3	5	16,3513
2	3	1	1	3	5	16,3690
3	1	0	1	3	5	16,3932
2	3	1	0	1	5	16,3996
3	3	0	1	3	5	16,4006
2	1	1	1	1	0	16,4059
3	3	1	1	1	0	16,4159

Tabla 65 Experimento de base de sintagmas nominales con codificación 2 del número de sílabas, con (5) o sin (0) codificación del signo de puntuación anterior, con (1) o sin (0) codificación de la posición de la sílaba en el final de una palabra, con (1) o sin (0) codificación de la posición de la palabra en la frase, con (1) o sin (0) codificación de la pertenencia a palabra función.

Si se realiza una búsqueda exhaustiva en el espacio de los últimos parámetros con los que hemos experimentado (Tabla 57), se encuentran algunas soluciones mejores, aunque no significativamente mejores.

4.7.6 Conclusiones sobre enunciativas

Los resultados obtenidos son coherentes con los anteriores, aunque ahora el tamaño del contexto debe ser ampliado al ser las frases más largas. La combinación del parámetro “signo de puntuación final” y “número de frase portadora” introduce mejoras significativas respecto a no considerarlos, aunque no por separado como ocurría en los experimentos de los nombres propios o las interrogativas.

Los nuevos parámetros ensayados apenas aportan mejoras.

La estrategia no exhaustiva se ha vuelto a mostrar como muy efectiva.

4.8 Experimentos con las frases especiales

Hemos dejado para el final las grabaciones que no eran a priori claramente asignables a una de las tres subdivisiones anteriores (nombres propios, interrogativas o sintagmas nominales), o bien que presentaban algún tipo de peculiaridad. Con estas frases haremos experimentos individuales y experimentos de agrupación con otras

frases.

4.8.1 Condiciones de experimentación

Siguiendo los experimentos de base previos, en los experimentos con las frases especiales emplearemos zscore, los elementos acentuada, inicial y final con tamaño de ventana del contexto variable, con codificación 2 del número de sílabas, usando 4 bits para codificar el signo de puntuación final del grupo fónico y con codificación del número de frase portadora.

4.8.2 Experimentos con las frases especiales 6 y 7

Las frases 6 y 7 son en cierta medida especiales: presentan la peculiaridad de que durante la grabación no se realizó la pausa en el momento adecuado (justo detrás de la palabra señor o señora, antes del apellido) y por necesidades del proyecto no fue posible volver a grabarlas. Dado que estas grabaciones están marcadas de manera incompleta, ha sido necesario parametrizarlas manualmente para tener en cuenta que el comienzo de grupo fónico que se marcó era ficticio en algunos casos. Realizando unos experimentos básicos, podemos observar que es mejor tratar las frases 6 y 7 como una unidad, en vez de como frase independientes, aunque de manera no significativa (17,54 Hz frente a 17,21). Experimentaremos con varios tamaños de capa oculta (5, 10 y 20), al ser menor el número de ejemplos.

4.8.2.1 Experimentos conjuntos con las frases 6 y 7

El número total de ejemplos de evaluación, usando *leave-one-out*, es de 850 sílabas y los mejores resultados son:

Tamaño del contexto	Neuronas Ocultas	Error Absoluto
2	5	17,2092
1	10	17,2233
1	20	17,2349

Tabla 66 Experimentos con las frases 6 y7.

4.8.2.2 Experimentos con la frase especial 6

El número total de ejemplos de evaluación, usando *leave-one-out*, es de 437 sílabas y los mejores resultados son:

Tamaño del contexto	Neuronas Ocultas	Error Absoluto
3	5	15,5310
3	10	15,6373
1	20	15,9741

Tabla 67 Experimentos con la frase 6.

4.8.2.3 Experimentos con la frase especial 7

El número total de ejemplos de evaluación, usando *leave-one-out*, es de 413 sílabas y los mejores resultados son:

Tamaño del contexto	Neuronas Ocultas	Error Absoluto
1	5	19,6413
1	10	19,6647
3	20	19,6761

Tabla 68 Experimentos con la frase 7.

4.8.2.4 Experimentos con las frases especiales 6 y 7 agrupadas con los demás nombres propios

El número total de ejemplos de evaluación, usando *leave-one-out*, es de 2536 sílabas y los mejores resultados son:

Tamaño del contexto	Neuronas ocultas	Error absoluto
3	10	13,6765
3	30	13,7123
4	20	13,7940
4	5	13,8719

Tabla 69 Experimento agrupando las frases 6 y 7 con los nombres propios.

4.8.2.5 Conclusiones sobre las frases 6 y 7

Si probamos a combinar estas frases con las de nombres propios (grupo al que deberían pertenecer), los resultados son peores que considerándolas un conjunto diferenciado, aunque de manera no significativa (12,84 Hz frente a 13,68). Esto confirma que se trata de grabaciones especiales que será necesario reconsiderar en el futuro.

4.8.3 Experimentos con la frase especial 8

La frase 8 también resulta especial, aunque por diferente causa: fue marcada por una persona distinta, en la fase final del proyecto y de manera incompleta (no todas las grabaciones fueron marcadas y segmentadas). Debido a posibles errores en los datos, no se incluyen experimentos con esta frase.

4.8.4 Experimentos con la frase especial 15

La frase 15 es especial porque se trata de un campo variable inserto dentro de una portadora interrogativa, pero no en posición final.

4.8.4.1 Experimentos con la frase especial 15 considerada como interrogativa

El número total de ejemplos de evaluación, usando *leave-one-out*, es de 2664 sílabas.

Tamaño del contexto	Neuronas ocultas	Error absoluto
3	10	13,7890
3	20	13,8222
2	30	13,8971
2	5	14,1634

Tabla 70 Experimento con la frase especial 15 considerada como interrogativa.

Apenas varían los resultados entre considerarla como interrogativa o no considerarla como tal (13,789 frente a 13,794).

4.8.4.2 Experimentos con la frase especial 15 considerada como enunciativa

El número total de ejemplos de evaluación, usando *leave-one-out*, es de 3272 sílabas.

Tamaño del contexto	Neuronas ocultas	Error absoluto
2	30	16,0920
2	10	16,1139
2	20	16,1957
3	5	16,4294

Tabla 71 Experimento con la frase especial 15 considerada como enunciativa

Apenas varían los resultados entre considerarla o no considerarla como integrante del grupo de las enunciativas (16,09 frente a 16,23), aunque es mejor agruparlas.

4.9 Experimento global conjunto con todas las frases

Tamaño del contexto	Neuronas ocultas	Error absoluto
3	30	15,0047
5	10	15,0659
5	20	15,0675
5	5	15,1723
4	20	15,1740
2	20	15,2159

Tabla 72 Experimento general conjunto con todas las frases

El número total de ejemplos de evaluación, usando *leave-one-out*, es de 8632 sílabas.

Aunque la diferencia no es significativa al 99% de confianza, es peor juntar todas las frases que agruparlas de la manera hasta ahora realizada (15,00 frente a 14,68).

4.10 Conclusiones sobre el modelado de F0 en dominio restringido

Los resultados obtenidos son en general coherentes con los resultados de (*J.A. Vallejo* 1998), aunque ahora ensayados en un dominio variado que va desde el habla aislada hasta los sintagmas nominales largos.

Al tratar grupos fónicos cortos, los parámetros relacionados con la posición de la sílaba en el grupo fónico resultan algo redundantes entre sí, pero no se ha encontrado una codificación mejor, y el número de sílabas no resulta muy importante.

El parámetro más importante, aquel que en general introduce diferencias significativas, resulta ser el “signo de puntuación final del grupo fónico”, que permite distinguir elementos variables con cadencias y elementos variables que generalmente presentaban anticadencias, y su incorporación es estadísticamente significativa.

El parámetro “número de frase portadora” introduce también mejoras significativas; parece que a pesar de que en las condiciones de grabación se intentó aislar el elemento variable de su frase portadora por medio de pausas obligatorias, no se consiguió. Es posible que la simple existencia de diferentes sesiones de grabación (que en términos generales coincidieron con las frases portadoras) haya podido introducir importantes diferencias en la curva de F0 de las distintas frases.

Los nuevos parámetros ensayados apenas aportan mejoras; tan sólo “es final de palabra”, “signo de puntuación inicial” y “es palabra función” mejoran algo pero nunca significativamente ni en todos los subdominios.

La estrategia de experimentación no exhaustiva (al ser comparada con la búsqueda exhaustiva del óptimo) ha mostrado su validez.

Es igualmente importante estudiar cómo agrupar las frases en subdominios, aunque las diferencias encontradas no han sido significativas y los perceptrones y los parámetros han mostrado gran robustez para asimilar subdominios diversos.

Capítulo 5 Análisis y síntesis de habla

con emociones

La variedad y la naturalidad van íntimamente unidas, siendo imprescindible la primera para conseguir la segunda; así, el habla sintética neutra, carente de matices emocionales, es doblemente artificial. Para poder incorporar variedad emocional a nuestra habla sintética, sea ésta por formantes o por concatenación, necesitaremos previamente realizar un análisis de habla emotiva, seguido por un modelado y una evaluación.

Dado que lo que pretendemos es extraer modelos para un sintetizador de voz que sea capaz de comunicar estados emocionales simulados, optaremos por la opción de grabar a un actor que lleve a cabo lo que queremos que realice el sintetizador: una simulación por medio de la voz. De esta voz no podremos, por tanto, extraer conclusiones sobre la verdadera naturaleza del habla emotiva, al no disponer de datos sobre auténtica voz triste, alegre o enfadada, pero servirá de modelo para que un sintetizador sea capaz de interpretar y transmitir como un actor.

5.1 Desarrollo de una nueva voz personalizable mediante síntesis por formantes

Como ya hemos señalado, una de las aplicaciones más importantes de la síntesis de voz con emociones la podemos encontrar en los comunicadores. En este contexto la síntesis de voz por formantes presenta una característica que la hace especialmente interesante: su flexibilidad. Al tratarse de un método de síntesis paramétrico, es posible personalizar una voz para un usuario concreto del comunicador.

El proyecto VAESS TIDE TP 1174 (*Voices Attitudes and Emotions in Synthetic Speech*) tenía por objetivo el desarrollo de uno de esos comunicadores portátiles para gente con discapacidad, que estuviese dotado de un sintetizador multilingüe, especialmente diseñado para ser capaz de comunicar no sólo las palabras y su significado, sino también actitudes y emociones que reflejen el estado del usuario.

Varias eran las voces sintéticas disponibles en castellano, basadas en síntesis por formantes (DECTALK, Tel-Eco o Rulsys-Infovox). Aunque la calidad inicial de este último era claramente menor, tres eran sus ventajas:

§ es multilingüe (no como Tel-Eco que es un sintetizador en castellano): esto permite dar mayor interés comercial al producto final (esto es, el comunicador).

§ teníamos acceso a la modificación interna de las reglas (al contrario que DECTALK): importante para conseguir mayor calidad en síntesis de voz con emociones.

§ La nueva versión de Infovox (GLOVE) cuenta con una nueva fuente glotal basada en el modelo de Fant; las experiencias previas lo muestran como ventajoso a la hora de implementar emociones (*I. Karlsson* 1994). Esta fuente glotal mejorada nos permitirá imitar mejor la similitud con una verdadera voz humana.

Entre las deficiencias del sintetizador elegido (detectadas en un proceso inicial de evaluación), destacaremos:

§ Una prosodia (duraciones y F0) con **acento marcadamente sudamericano**, poco adecuada para un producto destinado al mercado español y europeo.

- Una entonación bastante simple y esquemática, que no contemplaba preguntas ni exclamaciones. El cálculo de F0 en diptongos y triptongos necesita una reformulación que evite las desagradables subidas de F0.
- Las duraciones no respondían a un acento castellano (sino más bien mejicano), debido principalmente a que la diferencia de duración entre vocales acentuadas y no acentuadas era grande.

§ Diversos **errores en la conversión grafema a fonema**, así como en la asignación de acento (todo ello, por medio de reglas).

§ **Ausencia de algunos fonemas** importantes en castellano, como las palatales /ñ/ y /L/, o las oclusivas sonoras.

§ **Implementación no castellana de numerosos fonemas**, especialmente las vibrantes y las oclusivas.

Con el conocimiento y la experiencia adquirida en el mantenimiento del sintetizador Tel-Eco, el autor de esta Tesis sólo necesitó un breve curso de formación en el manejo de las herramientas propias de Rulsys, especialmente de su sistema de reglas dependientes del contexto.

Resumiendo, para desarrollar una voz con emociones, fue necesario implementar primero una nueva voz masculina por formantes de calidad, y para ello se decidió emplear el sintetizador de Infovox, caracterizado por disponer de un modelo paralelo de 5 formantes, una fuente glotal de *Liejencratz-Fant (I. Karlsson* 1994*)* y dotado de un sistema de reglas explícitas externas al intérprete, siguiendo la ya clásica división entre datos y procesos que resulta tan importante en varios campos del procesamiento natural del lenguaje.

5.2 Evaluación de la voz personalizada y del proceso de personalización

Como base para el desarrollo de una voz con emociones, ha sido necesario rediseñar completamente la voz sintética basada en formantes del sintetizador de *Infovox*. Como complemento al trabajo realizado en la consecución de la voz emotiva, se trabajó en que la voz resultase configurable, llegándose a implementar las emociones como particulares configuraciones personales de voz. Los detalles de este desarrollo pueden encontrarse en el apéndice A.3.1 “Personalización de voz”.

A continuación describimos el proceso final de evaluación de los resultados obtenidos.

5.2.1 Descripción de las sesiones de trabajo para la evaluación del proceso de personalización

En la evaluación del proceso de personalización de la voz han participado cinco personas. Todas ellos eran hombres, dado que la voz sintética disponible era masculina y no era realista que las mujeres intentasen personalizar una voz no femenina con la cual nunca se podrían sentir identificadas.

Cada sujeto se sometió a una sesión de al menos una hora, en la que debía partir de una voz estándar neutra como la desarrollada en la presente tesis, y cambiar los parámetros de configuración y personalización que se les ofrecían para conseguir personalizar la voz sintética. A lo largo de este proceso recibían asesoría de un experto que les explicaba el significado y posibles efectos de los cambios en los parámetros que el usuario podía realizar o se disponía a hacerlo (*S. Palazuelos et al 19*97). La labor del experto fue bastante dificultosa, puesto que los usuarios querían modificar parámetros que en lugar de ser sencillos y físicos (como eran los disponibles), fuesen complejos y perceptuales (edad de la voz, claridad, etc.). Como detalle de calidad de evaluación podemos señalar que uno de los participantes, cuando comenzó a notar que se estaba acostumbrando a la voz sintética, pidió interrumpir el proceso y recomenzarlo más adelante a fin de volver a aumentar su capacidad crítica.

El significado de los parámetros es el siguiente (entre paréntesis aparecen los valores adoptados para la voz estándar):

§ **Duración** (160): controla la velocidad de elocución.

§ **Onset** (100): regula el volumen y el arranque vocálicos.

§ **Rango** (100): diferencia entre los valores máximo y mínimo de la curva de tono fundamental F0.

§ **F0** (100): nivel medio de F0.

§ **Pendiente** (110): regula la pendiente media del contorno de F0 de cada frase.

§ **Tilt** (100): Mayores valores de este parámetro incrementan la presencia de altas frecuencias en la voz, con el consiguiente aumento de claridad y riqueza espectral.

§ **NA** (0): Ruido asociado a la fuente glotal y síncrono con ella.

§ **Ratio** (100): Relación entre la duración de las vocales acentuadas y no acentuadas. Permite cambiar asimétricamente la velocidad de elocución y articulación.

§ **OQ** (100): (*Open Quotient*) Relación entre el tiempo que la glotis permanece abierta y el tiempo que permanece cerrada, durante un periodo de tono.

§ **Emoción** (neutral): permite aplicar reglas específicas para la voz alegre y la voz neutra, cambiando la entonación resultante.

Los valores finales personalizados de los parámetros se compararon con la voz estándar y se evaluaron los problemas encontrados en el proceso.

5.2.2 Resultados

5.2.2.1 Valores personalizados de los parámetros para cada usuario

Usuario	Velocidad de elocución	Onset	Rango	F0	Pendiente	Tilt	NA	Ratio	OQ	Emoción
1	160	99	252	100	110	235	0	100	100	Alegre
2	150	160	150	100	120	170	85	125	100	Neutral
3	140	100	170	150	110	100	60	100	100	Alegre
4	120	60	90	100	110	100	0	125	100	Neutral
5	130	100	140	110	140	180	0	110	180	Neutral

Tabla 73 Valores de los parámetros para cada usuario.

Como resultado de las sesiones de trabajo los resultados son (*S. Palazuelos et al* 1997):

§ Dos de las personas fueron capaces de adaptar la voz a sus preferencias.

§ Otras dos personas consiguieron nuevas voces que alcanzaron una calidad aceptable (como lo era la voz estándar) y algunos de los rasgos deseados se hallaban presentes en dichas voces personalizadas, aunque tras una hora de cambios no obtuvieron la voz deseada y se quedaron con la última que habían conseguido. En todo caso, las preferencias dependían de la frase con la que se evaluase.

§ La persona restante no tuvo éxito en el proceso, obteniendo una voz que ni era preferible ni tan siquiera aceptable, posiblemente porque buscaba una voz de niño para la que el sistema no está diseñado.

Una característica común a todos los participantes fue su deseo de disponer de una voz más lenta que la estándar. Uno de los usuarios prefirió que la voz fuese más seria y profunda, lo que le obligó a reducir además el rango aunque se mantuvo dentro lo que podíamos catalogar como una voz neutra (no triste).

Otros rasgos comunes deseados fueron la expresividad y la juventud (la entonación se deseaba más dinámica, por lo cual incrementaron el rango, aunque el valor medio, por lo general, no se vea modificado); incluso se llegó a intentar que la voz fuese más alegre, menos monótona, haciendo que dos de los participantes manipulasen el parámetro emoción, (lo cual no estaba dentro del guión inicial) y se quedasen finalmente con una voz que incluía algunas reglas propias de la entonación alegre, confirmando parcialmente la conexión entre personalización, naturalidad y emotividad. Para algunos usuarios sería necesario poder adaptar la entonación según las frases, para hacerla más inteligente. Algunos cambios que sentaban bien a una frase, resultaban perjudiciales al cambiar de frase.

Un comentario general fue que el proceso resultaba largo y lento, con numerosos parámetros que controlar. La presencia del experto se hizo imprescindible para clarificar aquellos puntos que resultaban más misteriosos o confusos, o para aconsejar cómo modificar los parámetros a fin de conseguir el efecto final deseado.

5.2.3 Evaluación de la calidad global de la voz sintética

El cuestionario general de calidad de voz con el que finalizaba la sesión realizada con voz sintética nos ha permitido evaluar la calidad de la voz (*S. Palazuelos et al 19*97) con 15 oyentes que escucharon 5 frases cada uno (pudiendo repetir la audición de cada una).

5.2.3.1 ¿Cómo de natural suena la voz?

La voz suena muy natural	0
La voz suena más bien natural	2
La voz suena más robótica que natural	13

5.2.3.2 ¿Cómo es de inteligible el habla?

Buena	3
Aceptable	11
Pobre	1

5.2.3.3 ¿Cómo calificaría la calidad de la voz?

Buena	3
Aceptable	11
Pobre	1

Se puede observar que ninguno de los oyentes clasifica o juzga la voz sintética como natural y que la mayoría la califica de robótica. Sin embargo, la calidad general y su inteligibilidad son consideradas como aceptables y fáciles de entender.

Respecto a la calidad de voz, la inteligibilidad no fue homogénea a lo largo de las emociones empleadas, reduciéndose la identificabilidad de algunos fonemas como los laterales en determinados contextos[17].

Aunque los usuarios comentaron que la identificación resultaba más problemática al principio (cuando no tenían identificados los patrones rítmicos y entonativos que se podían dar), las frases largas y las frases alegres y enfadadas (como veremos posteriormente) resultaban más difíciles que el resto, posiblemente debido a que su modelado prosódico se revelaba como más irregular.

5.3 La base de datos SES: *Spanish Emotional Speech*

El primer paso necesario para producir síntesis de voz con emociones es el análisis de voz que simula emociones, etapa que nos permitirá posteriormente realizar un modelado de la misma que formará parte del sintetizador. Para ello precisamos grabar una base de datos de voz simulando estados emotivos (dado que buscamos realizar síntesis de voz, es preferible emplear un único actor, convenientemente evaluado, pues nuestro objetivo no es analizar y modelar la voz espontánea en esos estados, sino modelar voz que transmita la impresión de que corresponde a estos estados, y esa es una misión típica de un actor).

Aunque la base de datos que podíamos grabar era necesariamente pequeña, un requisito era que contuviera suficiente diversidad de fenómenos fonéticos y prosódicos para permitir cubrir un buen análisis y modelado del habla emotiva (*E.V. Enríquez et al 1996*). Con el fin de minimizar el efecto que un posible contenido semántica emotivo pueda tener sobre el habla, emplearemos textos preferiblemente neutros desde un punto de vista del contenido (aunque haya una excepción en el párrafo número 2). Agruparemos estos textos (se pueden consultar en el apéndice A.3.3 “Textos de la base de datos SES”) en 3 categorías (frases cortas, palabras y párrafos) interrelacionadas entre ellas, como a continuación describiremos.

5.3.1 Frases cortas

Como se ha dicho, se pretende, en primer lugar, que sean frases de carácter neutro, es decir, que no estén, preferentemente marcadas por ningún tipo de emotividad. Se han evitado, pues, verbos con un significado emotivo, así como las funciones conativas y la segunda persona. Mayoritariamente, las frases son en tercera persona; se han incorporado, sin embargo, algunas frases de primera persona que, sin embargo, son plenamente declarativas. Es de esperar que si a los significados meramente denotativos (de contenido semántico no emotivo)

se les incorporan otros valores expresivos (una voz que simula segmental y suprasegmentalmente una emoción), la identificación de la emoción transmitida haya que achacársela a los parámetros segmentales y suprasegmentales de la voz.

Se han confeccionado quince frases entre las que aparecen todos los fonemas del español, así como sus alófonos más representativos. Sin embargo, debido al propósito del proyecto para el que se grabó la base de datos (era el proyecto VAESS donde se empleaba síntesis por formantes), pareció oportuno no incluir un número excesivo de oclusivas, en especial, sordas (por su falta de energía), ni tampoco nasales (por la dificultad de aislar el formante nasal en las bajas frecuencias). Por otra parte, se buscó un mayor número de fricativas, líquidas (laterales y vibrantes) y realizaciones fricativas de consonantes oclusivas y de las vibrantes. Se han incluido casos, además, de los grupos consonánticos más habituales, incluyendo los favorecedores del elemento esvarabático.

La longitud de las quince frases cortas oscila entre las ocho y trece sílabas, con un mínimo de tres sílabas tónicas y un máximo de cuatro. Las palabras finales son, como es habitual en español, mayoritariamente paroxítonas, aunque se incluyen también dos terminaciones oxítonas.

Aunque se han incluido cinco estructuras de carácter interrogativo, la base de datos se centra en las declarativas, por lo que no se ha considerado incorporar el modelo entonativo de todas las posibles interrogativas del español, y que se ha evitado, además, el uso de los pronombres interrogativos.

5.3.2 Palabras aisladas

De las frases propuestas, se han entresacado treinta y una palabras aisladas (entendiendo como tales no sólo el concepto gráfico, sino también la cohesión e inseparabilidad prosódica con el determinante). Esto se ha hecho así, en primer lugar, porque en español es extraño el uso de determinados sustantivos aislados, sin determinantes, y el considerarlos así, en algunos casos, podría forzar en exceso su pronunciación. Dado que se buscará naturalidad dentro de los límites impuestos por la simulación actoral, con algunos términos se prefirió mantener la unión sintagmática. En segundo lugar, debemos tener presente que el considerar un elemento como palabra aislada es meramente histórico y gráfico, en especial en los clíticos verbales, por lo que también en estos casos hemos preferido incluir algunos grupos acentuales tal y como aparecían en la oración original (así se grabará “el final” o “se cayó” en vez de “final” o “cayó”). Por último, sólo en un caso (1ª oración) hemos considerado un grupo sintagmático no tan íntimamente cohesionado (adverbio + verbo); sin embargo, en este caso lo que se ha pretendido mantener es el matiz negativo del enunciado (“no queda”).

Dentro de este grupo se favorece, además, la comparación entre las palabras dentro de la secuencia oracional (en posición inicial, interior o final) y su pronunciación aislada, por lo que es probable que se puedan establecer diferencias acusadas entre unas y otras.

5.3.3 Párrafos de corta longitud

Incorporar la lectura de párrafos cortos puede aportar alguna luz a la hora de establecer diferencias entre los distintos modelos entonativos, en especial, en las diferentes estructuras sintácticas. De ahí que se hayan considerado dentro de la base de datos tres párrafos de entre cuatro y ocho líneas, de carácter neutro y donde, como en el caso de las frases cortas, se ha evitado el uso de verbos y estructuras de marcada emotividad. Además, se ha incorporado un cuarto párrafo, en el que se incorporan, en el marco de una breve estructura narrativa, doce de las quince frases cortas. Evidentemente, esto facilitará información en cuanto a las diferencias que puedan observarse, no sólo en función de los diferentes modelos emocionales entonativos, sino comparar un mismo modelo en tres contextos diferentes.

5.3.4 Grabación

Los textos que se acaban de describir fueron grabados en una sala acústicamente aislada usando un micrófono de mesa de alta calidad, una tarjeta de sonido Oros con programa Europec y una grabadora digital Sony. La frecuencia de muestreo fue de 16 Khz. Los textos fueron interpretados en 3 ocasiones[18] cada uno por un actor profesional de 38 años, con acento castellano y con más diez años de experiencia. Aunque en el proyecto VAESS sólo se emplearon las emociones primarias simuladas (tristeza, alegría y enfado), se grabó igualmente una emoción calificada como secundaria (sorpresa) que pudiese ser empleada posteriormente.

Los textos fueron proporcionados al actor con anterioridad al día de grabación, dado que no se buscaba la espontaneidad, sino la mejor interpretación posible. La múltiple grabación posibilitó, sin embargo, que el actor interpretase una misma frase de distintas maneras conscientes, empleando patrones entonativos y rítmicos diferentes. Todo ello fue realizado por el actor sin someterse a ningún esquema o modelo prefijado, siendo libre de decidir cómo debía simular los estados emocionales que se le encomendaron.

5.3.5 Etiquetado y marcado de SES

Los dos mil fonemas por emoción fueron etiquetados fonéticamente de manera completamente manual con la ayuda de la herramienta de edición de voz PCV desarrollada en el proyecto VAESS como antes se mencionó.

El marcado de F0 se llevó a cabo semiautomáticamente, con el marcador de periodos de la frecuencia fundamental, empleado en síntesis por difonemas, adaptado para procesar elocuciones más largas. Los resultados obtenidos de esta manera fueron visualmente revisados usando el mismo programa. Finalmente se resintetizaron las grabaciones por medio de concatenación de difonemas, linealizando la curva de F0 en el nivel de sílaba y cuantificando las duraciones a un número entero de periodos de F0. Por ello fue necesario emplear un algoritmo de concatenación con modificaciones prosódicas menores durante la resíntesis. Esta segunda revisión por resíntesis se tradujo en nuevas correcciones, especialmente de la curva de F0.

5.3.6 Análisis de SES

Ilustración 2 Fragmento de voz neutra (parte superior) y su correspondiente de voz enfadada (parte inferior).

5.3.6.1 Análisis cualitativo

Una simple audición y una inspección visual de la forma de onda del material grabado revela ya las principales características que pueden ser las que ayuden a su identificabilidad:

§ **Enfado**: es, posiblemente, la emoción más destacada. Se aleja de lo comúnmente señalado en la bibliografía, debido a que se trata, no de un enfado típico en caliente (caracterizable con su notable intensidad y tono medio), sino más bien de una amenaza verbal que sugiere un enfado en frío. Observando la forma de onda, apreciamos un notable esfuerzo vocal que se traduce en la presencia de una distorsión que la hace muy característica. En la ilustración se puede apreciar la notable diferencia entre un fragmento de voz neutra y un fragmento equivalente de voz enfadada, donde se aprecian las

resonancias e irregularidades de tono que distinguen la simulación de enfado que ha hecho nuestro actor, y que han dificultado notablemente la labor de marcado.

§ **Tristeza**: se aprecia una cierta monotonía de tono (lógica en un estado de baja excitación), una velocidad de elocución lenta (pero con una velocidad de articulación que no se desvía mucho de la neutra); también se aprecia una cualidad lánguida en la voz, con suspiros en las pausas, más abundantes y más largas. La intensidad de estos ficheros es menor, pero al haberse empleado un micrófono de mesa, no se puede concluir esto directamente de las grabaciones, aunque el actor sólo varió su situación en la sala de grabación para la alegría.

§ **Alegría**: desde un punto de vista prosódico destaca la presencia de varios patrones entonativos en las frases de modalidad enunciativa, variando la posición del foco de una manera no sistemática, pudiéndose dar el caso de que, para un mismo texto pero en 2 sesiones diferentes, la palabra realzada pasa del principio al fin de la frase. Subjetivamente se puede apreciar que si la frase no comenzó con suficiente énfasis prosódico (ritmo y entonación altos, propios de un estado de excitación), entonces la parte final se ve enfatizada, resultando en una curva de declinación ascendente en lugar de descendente. Por el contrario, si el comienzo fue demasiado enfático, la parte final lo compensa reduciendo apreciablemente el nivel de F0, dando lugar a una declinación estándar descendente, aunque con una pendiente mayor de lo habitual. En otras ocasiones el actor puede seleccionar el foco en mitad de la frase, debido a la presencia de una palabra en torno a la cual el actor decide centrar la frase (por ejemplo, la palabra "mucho" en "le gusta mucho el gregoriano"[19] o la palabra "deuda" en "dejaron la deuda al cero"). También desde un punto de vista subjetivo, la voz alegre se caracteriza por la presencia de voz emitida sonriendo y por una claridad y brillantez notables.

§ **Sorpresa**: estas grabaciones son únicas desde un punto de vista entonativo, puesto que su tono medio es muy elevado, acentuado por la presencia de una gran anticadencia final que alcanza unos niveles de F0 propios de una voz de mujer o un falsete (>250 Hz.).

5.3.6.2 Análisis cuantitativo de las duraciones y el ritmo

Para analizar y comparar los datos entre las distintas emociones emplearemos un modelo multiplicativo. La solución del sistema de ecuaciones la calcularemos en un dominio logarítmico (donde las ecuaciones se convierten en un sistema lineal sobredimensionado con más ecuaciones que incógnitas), para luego volver al dominio lineal y realizar la búsqueda de un óptimo local (*Levenberg-Marquardt*). Los detalles sobre este proceso pueden consultarse en (*G. Martínez Salas* 1998). Hemos de destacar que el efecto del alargamiento prepausa se divide en 3 factores prepausa (alargamiento prepausa de vocal, alargamiento prepausa de fonema no continuo y alargamiento prepausa de fonema continuo) y se ha distinguido entre monosílabos con y sin acento.

El error medio de modelado está en torno al 20%, excepto para el enfado, cuyo error supera el 25%. Esto puede estar relacionado con las dificultades de etiquetado y marcado debidas al ruido y distorsión que caracteriza la simulación de voz de nuestro actor.

Para el análisis de la variación de los parámetros entre las distintas emociones, tomaremos la voz neutra como referencia, y agruparemos los 123 parámetros en varios grupos: el factor multiplicativo medio aplicado a las vocales, a las consonantes y a los diptongos (factor medio por el que se han multiplicado); el factor medio de alargamiento prepausa; el factor medio debido al número de sílabas; la duración media de los todos fonemas, de las vocales, de las consonantes y de los diptongos.

	Alegría / Neutra	**Tristeza / Neutra**	**Sorpresa / Neutra**	**Enfado / Neutra**
Efecto medio del contexto para las consonantes	0,9222	1,0607	1,0224	0,9831
Efecto medio del contexto para los diptongos	0,9627	1,1620	1,1233	1,0538
Efecto medio del contexto para las vocales	0,9969	1,1003	1,1067	0,9168
Efecto del alargamiento vocálico prepausa	0,9086	1,2816	1,0398	0,7811
Efecto medio del número de sílabas	1,0116	1,1326	1,1903	1,0961

Duración media de todos los fonemas	1,0498	1,2629	1,1464	1,2289
Duración media de las vocales	1,0664	1,0296	1,1164	1,1003
Duración media de las diptongos	0,9952	1,0736	1,0622	1,0208
Duración media de las consonantes	1,1114	1,5303	1,2516	1,4994

Tabla 74 Variación de diversos parámetros de duración entre las distintas emociones.

Podemos observar que la voz triste es la más lenta, aunque no mucho más que el enfado, y que el alargamiento prepausa de las vocales finales es mucho mayor en la tristeza.

El análisis de los párrafos no se realizó completamente. La prosodia de los párrafos no se tuvo en cuenta debido a que en un experimento sobre la capacidad de la prosodia enfadada para transmitir ese estado emocional, esta capacidad se reveló como muy baja (la influencia de la prosodia en la identificación del enfado simulado por el actor era casi nula, como veremos más adelante).

Al comparar el modelado de las frases con el modelado de los párrafos, podemos constatar que las emociones no se caracterizan por unas duraciones absolutas constantes, ya que en todo caso la velocidad de elocución es mayor en los párrafos que en las frases, con independencia de la emoción simulada. Esto es consistente con lo observado en las bases de datos de voz neutra, lo cual confirma la generalidad del fenómeno.

Frases / párrafos	**Neutra**	**Alegría**	**Tristeza**	**Sorpresa**
Efecto medio del número de sílabas	1,0582	1,0794	1,1025	1,2068
Duración media	1,1442	1,1894	1,2837	1,2445
Alargamiento vocálico prepausa	1,0963	1.2543	1,0876	1,0344

Tabla 75 Ratio entre el modelado de duración en las frases y en los párrafos para las distintas emociones.

Signo de puntuación	**Neutra**	**Alegría**	**Tristeza**	**Sorpresa**
Duración media para el punto	0,910	0,420	1,176	0,547
Desviación típica para el punto.	0,167	0,070	0,182	0,075
Duración media para otros signos	0,514	0,316	0,697	0,346
Desviación típica para otros signos	0,137	0,082	0,131	0,074

Tabla 76 Duración media de las pausas por signos de puntuación en las distintas emociones.

De acuerdo con nuestras intuiciones, la tristeza presenta mayor duración de la pausas (menor velocidad de elocución), y por el contrario, la alegría (una emoción con mayor actividad) presenta la menor duración.

5.3.6.3 Análisis cuantitativo de la entonación

Para el análisis de la entonación emplearemos un modelo de picos y valles, dividiendo la frase en 3 zonas: hasta la primera tónica, entre la primera tónica y la última, y a partir de la última tónica. Entre pico y pico estableceremos una recta de declinación y lo mismo entre los valles (vocales inmediatamente anteriores a cada tónica que no formen con ella un diptongo o un hiato). La zona inicial tendrá 3 parámetros característicos (valor de F0 en la primera sonora, en la vocal anterior a la primera tónica y el valor en la primera tónica), mientras que la zona final se caracteriza por otros 3 parámetros (la última tónica, el último valle y el último fonema). Todos estos parámetros serán estimados de la base de datos mediante regresión lineal, aunque se distingue entre terminaciones oxítonas y no oxítonas, y entre oraciones interrogativas y enunciativas. Los detalles pueden verse en (*G. Martínez-Salas* 1998). En la siguiente tabla comparamos los resultados obtenidos para las frases:

	Alegría/ Neutra	**Tristeza/ Neutra**	**Sorpresa/ Neutra**	**Enfado/ Neutra**
F0 de la primera tónica	1,29	0,83	1,61	0,96
Pendiente de declinación de las tónicas	1,82	0,76	-1,44	-0,05
F0 de la 1ª sílaba	1,23	0,76	1,12	0,90

F0 del último valle no oxítono (enunciativa)	0,91	0,68	1,47	0,90
F0 de la última tónica no oxítona (enunciativa)	1,32	0,79	2,51	1,19
F0 del último fonema no oxítono (enunciativa)	1,07	1,00	1,78	1,25
F0 de la 1ª sílaba (interrogación)	1,08	0,76	1,06	0,95
F0 del último valle (interrogativa)	1,15	0,84	1,45	1,13
F0 e la última tónica (interrogativa)	1,55	0,91	1,18	1,34
F0 del último fonema (interrogativa)	1,12	0,64	1,56	0,90

Tabla 77 Resultados del análisis cuantitativo de la entonación de las frases para las diversas emociones.

Podemos describir cada una de las emociones de la siguiente manera:

§ **Alegría**: su tono inicial es superior al neutro, más por el valor de su primera tónica que por su sílaba inicial. La declinación, sin embargo, es superior, aunque en la última tónica también es mayor el valor cuando el actor simula estar alegre. El tono final es similar y no parece que transmita emoción alguna, lo mismo que los valles o su pendiente de declinación. Las oraciones interrogativas presentan un tono final más elevado, aunque la última tónica es la que alcanza el valor más elevado.

§ **Tristeza**: sus valores de F0 son inferiores a la voz neutra, tanto en la primera tónica como en la última, por lo cual presenta un rango y una pendiente de declinación similares. En las oraciones interrogativas el fonema final presenta menor F0, como suele ser en una emoción poco activa. Los valles presentan menor F0, algo necesario para que los picos sean percibidos como tales, y que contribuyen a la impresión general de baja actividad. Las menores desviaciones típicas nos indican una voz muy homogénea y repetitiva.

§ **Sorpresa**: sus niveles de F0 son los mayores que hemos observado, tanto en la primera tónica, como en la última, presentando una pendiente de picos negativa, esto es, creciente. Las interrogativas con sorpresa se caracterizan por una muy fuerte subida entre la última tónica y el último fonema, siendo muy relevante la diferencia de valores entre las últimas tónicas cuando se enuncia y cuando se pregunta. Los valles, que no parecen ser portadores de información emocional, son similares a los de la voz neutra, contrastando importantemente con los picos, lo cual produce un gran rango.

§ **Enfado**: presenta un tono muy plano, sin apenas declinación, con valores no muy alejados de la voz neutra. Las interrogativas se diferencian, sin embargo, por no tener su valor máximo en el último fonema, sino en la última tónica.

	Alegría/ Neutra	**Tristeza/ Neutra**	**Sorpresa/ Neutra**
F0 de la primera tónica	1,16	0,76	1,67
Pendiente de la declinación de las tónicas	1,31	0,57	-1,09
F0 de la 1ª sílaba	1,03	0,74	1,27
F0 de la última tónica no oxítona (punto)	1,21	0,88	2,71
F0 del último fonema no oxítono (punto)	0,95	0,82	2,43
F0 de la última tónica (coma)	1,5	0,91	2,42
F0 del último fonema (coma)	1,09	0,59	1,76

Tabla 78 Resultados del análisis cuantitativo de la entonación de los párrafos para las diversas emociones.

El análisis de los párrafos es similar, aunque no se analiza el enfado por los motivos expuestos anteriormente.

§ **Alegría**: resulta menos marcada en cuanto a su curva de F0 (presenta menor F0 inicial, aunque lo compensa subiendo la F0 en la zona final, y subiendo el valor de *continuation rise* antes de pausa no final de frase).

§ **Sorpresa**: también varía ligeramente su perfil, subiendo el tono inicial y disminuyendo la subida final, aunque sigue presentando declinación negativa.

§ **Tristeza**: se mantienen las tendencias observadas en las frases, con una *continuation rise* pequeña, y con ligero incremento de la monotonía (tono y rango algo menores).

5.3.6.4 Síntesis por formantes de voz con emociones

La incorporación de las diversas emociones seguirá 2 vías: su incorporación como si se tratase de una personalización de voz (definiendo qué valores adoptan los parámetros personalizables en cada emoción) y la creación de alguna regla específica para la síntesis de emociones (como, por ejemplo, importantes incrementos de F0, se deben traducir en anchos de banda mayores).

Aunque la mejora en la fuente glotal ha supuesto una clara mejora en la cálida brillantez de la voz alegre, no se ha podido caracterizar la voz de enfado en frío, amenazante, que el actor empleó, obligando a implementar un enfado caliente totalmente diferente, un enfado más parecido a los documentados en la bibliografía.

5.4 Evaluación del habla con emociones empleando síntesis por formantes

La evaluación de la calidad de la voz sintética desarrollada en el proyecto VAESS, debido a estar destinada a formar parte de un comunicador orientado a personas con discapacidad, debe comprender dos aspectos fundamentales: la calidad del habla sintética emotiva y la capacidad de personalización de la voz (*S. Palazuelos et al 1997*).

En cuanto a la calidad del habla con emociones, debemos evaluar el grado de identificabilidad de las diferentes emociones o tipos de voz que incorpora el sistema, esto es, la tristeza, el enfado, la alegría y la voz neutra. Dado que en el proyecto se buscaba el desarrollo de un comunicador personal, es necesario evaluar la flexibilidad del sistema para conseguir que la voz sintética empleada pueda adaptarse a las necesidades o gustos de un usuario que, por lo general, no será experto en materias de tecnología del habla.

5.4.1 Parámetros generales de la evaluación

En la evaluación de la calidad de voz han intervenido 17 sujetos que reunían las características de ser oyentes normales, hombres y mujeres, cuyas edades se encontraban comprendidas entre 18 y 55 años, provenientes de diferentes entornos sociales, que no se encontraban habituados a escuchar voz sintética. Catorce personas participaron sólo en una de las pruebas, mientras que otras tres tomaron parte en ambas; de esta manera el *test* de voz con emociones incluyó a quince personas, y el de personalización de la voz a tan sólo cinco. Este desequilibrio entre estas dos pruebas viene motivado por la diferente duración e implicación de ambas, mayor en el segundo caso.

5.4.1.1 Estímulos

A fin de no cansar en exceso a los sujetos participantes, y teniendo en cuenta que se deben evaluar cuatro tipos de estímulos (uno por cada tipo de emoción), hemos limitado a cinco los textos semánticamente neutros de las frases que evaluar. Dichas frases son las siguientes:

1) No queda fruta los viernes.
2) El final del siglo veinte.
3) Tengo la llave en el bolsillo.
4) Los participantes en el congreso marcharon después a El Escorial.
5) Se ganaba la vida vendiendo recuerdos alusivos a la Virgen Morenita, desde llaveros a platos con la imagen grabada en esmalte vidriado.

Se ha procurado mezclar frases extraídas de las grabaciones de párrafos disponibles (la cuarta y la quinta), con frases breves como las que también contiene la base de datos. Por su brevedad, no se ha considerado necesario incluir palabras aisladas, aunque pueden tener su importancia desde un punto de vista comunicativo.

5.4.2 Sesiones de trabajo con los oyentes

El trabajo con los sujetos del *test* de evaluación se ha realizado en dos sesiones, una para voz sintética y otra para voz natural del actor, evitando mezclar elocuciones de ambos tipos en la misma sesión. En cada sesión se les

presentaba una a una las veinte elocuciones de voz masculina a los oyentes: 5 frases simulando 4 emociones, presentadas en orden aleatorio, cuya emoción simulada o transmitida debía identificar el oyente. Aunque la elección se basaba en una lista cerrada (triste, alegre, enfadada y neutra), se han incluido dos opciones de naturaleza más abierta (emoción no reconocida y otras emociones), para contemplar los casos en que el oyente no fuese capaz de identificar o asociar ninguna emoción a la grabación que se le presentaba, o bien identificase una emoción no presente en la lista. Conforme va escuchando las elocuciones, el oyente debe rellenar el cuestionario de identificación En el caso de la prueba de voz sintética, una vez realizada, los participantes debían rellenar un cuestionario general de calidad sobre naturalidad e inteligibilidad. Ambos cuestionarios se pueden consultar en los apéndices A.3.4 “Cuestionario de evaluación de voz emotiva en el proyecto VAESS” y A.3.5 ”Cuestionario sobre la personalización de voz”.

5.4.3 Resultados

A continuación presentamos las tablas de resultados totales y parciales obtenidos. En negrita se ha señalado la columna con el número de aciertos para cada grabación (cada fila). Los números indican el número de personas que eligieron esa emoción (columna) para esa grabación (fila).

5.4.3.1 Identificación de la emoción transmitida por la voz sintética

En esta primera prueba deseamos comprobar la capacidad que tiene la voz sintética para transmitir un determinado estado emocional a los sujetos de la evaluación.

Grabación	**Emoción simulada**	**Identifican el habla como...**						**Número de frase**
		neutral	**alegre**	**triste**	**enfadada**	**No identifican**	**otras emociones**	
1	**Alegre**	4	2	2	1	5	Asombro	1
2	**Enfadada**	3	1	0	8	2	Afirmación	1
3	**Neutra**	8	5	2	0	0	0	4
4	**Triste**	8	0	6	1	0	0	5
5	**Triste**	0	0	13	0	0	Lento Desanimado	2
6	**Alegre**	4	6	0	3	1	Afirmación	4
7	**Neutra**	9	1	1	3	0	Afirmación	2
8	**Alegre**	4	7	3	0	1	0	5
9	**Neutra**	4	0	11	0	0	0	2
10	**Triste**	0	0	13	1	1	0	3
11	**Enfadada**	0	5	0	9	0	Afirmación	3
12	**Triste**	5	0	9	1	0	0	4
13	**Neutra**	11	0	4	0	0	0	3
14	**Alegre**	0	13	0	0	0	Afirmación Euforia	3
15	**Triste**	0	0	14	1	0	0	1
16	**Alegre**	2	8	0	1	4	0	2
17	**Neutra**	11	0	0	4	0	0	4
18	**Neutra**	10	0	1	3	1	0	1
19	**Enfadada**	0	3	0	9	1	Euforia Stress	5
20	**Triste**	2	0	13	0	0	0	5

Tabla 79 Resultados de la evaluación de identificación de la emoción transmitida por la voz sintética.

5.4.3.2 Matrices de confusión para voz sintética

Muestran el número de veces que el usuario ha identificado una elección cuando las grabaciones simulaban esa misma u otra. Los resultados ideales supondrían una matriz diagonal, dado que los errores están contenidos en las casillas fuera de dicha diagonal. Las filas recogen los resultados de reconocimiento al sintetizar voz con una determinada emoción. Las columnas muestran cuál era la emoción simulada, cuando los participantes en la evaluación reconocieron una determinada emoción (acertando o no).

5.4.3.3 Resultados totales de reconocimiento de la emoción simulada

Sintetizada como...	Identificada como ...					
	neutra	alegre	triste	Enfadada	no identifican	otra
Neutral	44 (**58,6%**)	0	22 (29,3%)	8 (10,6%)	1 (1,3%)	0
Alegre	18 (24%)	35 (**46,6%**)	7 (9,3%)	2 (2,6%)	10 (13,3%)	3 (4%)
triste	7 (9,3%)	0	62 (**82,6%**)	3 (3,9%)	1 (1,3%)	2 (2,6%)
enfadada	16 (21,3%)	16 (21,3%)	1 (1,3%)	32 (**42,6%**)	4 (5,3%)	6 (8%)

Tabla 80 Resultados totales de identificación de la emoción simulada.

5.4.3.4 Resultados para las 10 primeras grabaciones

Sintetizada como...	Identificada como...					
	neutra	alegre	triste	enfadada	no identificada	otra
Neutra	12 (**40%**)	0	17 (56,6%)	1 (3%)	0	0
Alegre	16 (36,3%)	14 (**31,8%**)	7 (15,9%)	1 (2,27%)	5 (11,4%)	Asombro 1 (2,27%)
Triste	0	0	26 (**86,6%**)	1 (3,3%)	1 (3,3%)	2 (6,6%)
enfadada	16(36,3%)	7(15,9%)	1(2,27%)	14(31,8%)	3(6,8%)	Afirmación 3 (6,8%)

Tabla 81 Resultados de identificación de la emoción simulada para las 10 primeras grabaciones.

5.4.3.5 Resultados para las 10 últimas grabaciones

Sintetizada como	Identificada como...					
	neutra	alegre	triste	enfadada	no identificada	otra
neutra	32 (**61,5%**)	0	4 (7,7%)	7 (13,5%)	1 (1,9%)	8 (15,3%)
alegre	0	13 (**86,6%**)	0	0	0	Afirmación 1 Euforia 1 (13,3%)
triste	7 (16,6%)		33 (**78,5%**)	2 (4,7%)	0	0
enfadada	0	8 (26,6%)	0	18 (**60%**)	1 (3,3%)	Euforia 1 afirmación 1 (10%)

Tabla 82 Resultados de identificación de la emoción simulada para las 10 últimas grabaciones.

Como se puede observar, los resultados van desde un 42 por ciento para el enfado, hasta el 82 por ciento en el caso de la tristeza. Dividiendo las sesiones en 2 fases (distinguiendo entre las 10 primeras grabaciones escuchadas y las 10 últimas), se observa cómo rápidamente los usuarios se acostumbran a la voz sintética y mejoran considerablemente los resultados de identificación. Inicialmente parece que sólo son capaces de identificar correctamente la tristeza pero, conforme avanza la prueba, los resultados suben desde un 31 – 40 por ciento hasta un 60 – 86 por ciento. Los comentarios de los participantes nos señalaban que tras escuchar las primeras grabaciones eran capaces de percibir las diferencias entre las emociones e identificarlas.

Sin embargo, la confusión remanente es todavía elevada para algunas emociones (por ejemplo, la voz enfadada se confunde con la alegre más de un 25 por ciento de las veces, a pesar de que las alegres son identificadas perfectamente al final). Algunos errores no son graves, como cuando se confunde la alegría con la euforia.

La emoción más fácilmente identificable siempre fue la tristeza, con más de un 86 por ciento de aciertos.

5.4.3.6 Resultados para voz natural

La referencia para evaluar nuestros resultados debe ser una prueba realizada con la voz natural del actor simulando las emociones consideradas. Los resultados obtenidos de esta manera mejoran considerablemente los de la voz sintética. La tabla muestra que los sujetos no tuvieron mayores problemas a la hora de identificar las emociones simuladas por el actor profesional; los números de la diagonal están claramente por encima del nivel de azar (20 por ciento). Un test Chi-*square* refuta por $p<0.05$ la hipótesis nula (que los resultados puedan haber sido obtenidos por selección aleatoria).

La emoción de más difícil identificación fue la alegría, que sólo dio un 74 por ciento de aciertos, confundiéndose con la voz neutra en más de un 17 por ciento de las ocasiones. Para el resto de las emociones, los resultados oscilan entre un 89 y un 90 por ciento.

5.4.3.7 Matrices de confusión para voz natural

Simulada como...	**Identificada como...**					
	neutra	**alegre**	**triste**	**enfadada**	**no identificada**	**otra**
Neutra	67 (**89,3%**)	1 (1,33%)	1 (1,33%)	3 (3,99%)	3 (3,99%)	
Alegre	13 (17,3%)	56 (**74,6%**)	1 (1,33%)	1 (1,33%)	3 (5,33%)	Stress
Triste	1 (1,33%)	0	70 (**90,3%**)	1 (1,33%)	2 (3,99%)	Melancolía
Enfadada	0	1 (1,33%)	2 (2,66%)	67 (**89,3%**)	4 (6,66%)	Énfasis

Tabla 83 Matriz de confusión para la voz natural.

5.4.3.8 Identificación de la emoción simulada en función del número de frase

Como los textos con los que hemos evaluado son de procedencia diversa, consideramos necesario evaluar si el texto o el tipo de texto influye significativamente sobre los resultados obtenidos.

Para la voz natural no son significativas las diferencias de acierto a la hora de clasificar grabaciones que provienen de párrafos o de las frases cortas, sugiriendo que, para las emociones consideradas (excluyendo la sorpresa), los párrafos y las frases pueden resultar igualmente identificables.

En cuanto a la voz sintética, las diferencias observadas en la tercera frase pueden deberse a la tristeza, que se hallaba presente 3 veces entre las 10 grabaciones no iniciales, y que goza de mayor tasa de reconocimiento por parte de los usuarios.

Número de la frase	**Identificaciones correctas para voz sintética**	**Identificaciones correctas para voz natural**
1	56,6%	86,6%
2	46,6%	86,6%
3	76,6%	93,3%
4	46,6%	93,3%
5	61,6%	83,3%

Tabla 84 Resultados de identificación para la voz sintética y para la voz natural.

5.5 Conclusiones sobre síntesis de voz con emociones mediante síntesis por formantes

Se ha creado un sistema completo de conversión texto a voz en castellano, empleando una nueva voz personalizable, y se han incorporado las emociones como una personalización más de la voz, todo ello empleando síntesis basada en formantes.

Tanto el proceso de personalización como la voz sintética con emociones han sido evaluado con usuarios satisfactoriamente. Los resultados de identificación de emoción a partir de la voz han resultado especialmente prometedores puesto que los humanos parecen adaptarse con rapidez a la voz sintética con emociones, y el periodo

de adaptación podría resultar breve, especialmente en el caso de la tristeza (siendo peor para el enfado). Es cierto que hubo problemas de inteligibilidad en algunos contextos fonéticos, pero los resultados son perfectamente aceptables, incluso si la voz evaluada no resulta totalmente natural. De acuerdo con los resultados de evaluación, la voz sintética con emociones desarrollada en el proyecto VAESS es comparable al estado de la cuestión en otros idiomas a nivel mundial. Aunque en los planteamientos iniciales del proyecto VAESS se consideró que eran independientes los módulos prosódico y segmental del sintetizador, la conclusión de nuestro trabajo dista mucho de ser esta: debemos señalar que las trayectorias de los formantes pueden provocar, al incrementarse la velocidad de elocución, ruidos de naturaleza seudo-oclusiva, que es necesario limar uno a uno, modificando la reglas segmentales.

Otro resultado importante fue la creación de la primera base de datos de habla emotiva simulada en castellano: Está orientada a síntesis prosódica y al análisis de la prosodia en párrafos y frase cortas mediante técnicas paramétricas, y dio lugar a un modelado diferencial de cada emoción respecto a la voz neutra.

5.6 Experimentos de síntesis-por-copia y voz con emociones

Tras los resultados prometedores en síntesis por formantes, pareció adecuado experimentar con la síntesis por concatenación de difonemas. Dado que nuestra base de datos estaba orientada a prosodia y no es suficientemente grande, nos limitaremos a experimentos de síntesis-por-copia (*copy-synthesis)*, de los que sin embargo podremos obtener algunos resultados que esperamos sean de interés general.

El primer experimento consistirá en validar nuestro sistema de síntesis-por-copia, el mismo que hemos empleado en la revisión del marcado de nuestra base de datos. 21 nuevas personas se sometieron a una nueva prueba de evaluación de carácter cerrado con la opción de "emoción no identificada". Para este primer experimento tomaremos tanto los difonemas como la prosodia de la voz natural aunque, como hemos indicado, la prosodia se verá linealizada en el nivel de sílaba. Esta distorsión o ruido puede, obviamente, afectar a la capacidad de identificación de los humanos, no tratándose, por tanto, de un experimento redundante respecto al de evaluación de voz natural. El modelo de prosodia así simplificado se corresponde con el método empleado por el sintetizador del GTH.

Los resultados de esta síntesis-por-copia, aunque se encuentran por encima del nivel de selección aleatoria según un test de la T de Student ($p>0.95$), se encuentran significativamente por debajo de las tasas de la voz natural (excepto para el enfado amenazante en frío, donde los resultados fueron mejores, aunque no significativamente, posiblemente porque las pequeñas distorsiones convirtieron la voz en más amenazante si cabe).

Sintetizada como...	**Identificada como...**				
	neutra	**alegre**	**triste**	**enfadada**	**no identificada**
neutra	76,2 %	3,2%	7,9%	6,3 %	4,8%
alegre	3,2%	61,9%	9,5%	7,9%	6,3%
triste	3,2%	0%	81,0%	0%	11,1%
enfadada	0%	0%	0%	95,2%	4,8%

Tabla 85 Resultados de identificación de emociones generadas mediante el método de síntesis por copia.

En un segundo experimento decidimos investigar hasta qué punto el ritmo y la entonación pueden servir a un humano para identificar la emoción expresada. Se tratará de un nuevo experimento de síntesis-por-copia o resíntesis, en el que los difonemas extraídos de una grabación donde el actor pretendía transmitir una determinada emoción, se ven resintetizados empleando las duraciones y el contorno de F0 de una grabación de la misma frase pero con diferente emoción simulada.

Prosodia copiada de...	Difonemas copiados de...	Identificada como...					
		neutra	alegre	triste	sorpresa	enfadada	otra
Alegre	**Neutra**	52.4 %	19%	11.9%	4.8%	0%	11.9%
Triste	**Neutra**	23.8%	0%	66.6%	0%	2.4%	7.1%
Sorprendida	**Neutra**	2.4%	16.7%	2.4%	76.2%	0%	2.4%
Enfadada	**Neutra**	11.9%	19%	19%	23.8%	7.1%	19%
Neutra	**Alegre**	4.8%	52.4%	0%	9.5%	26.2%	7.1%
Neutra	**Triste**	26.2%	2.4%	45.2%	4.8%	0%	21.4%
Neutra	**Sorprendida**	19.0%	11.9%	21.4%	9.5%	4.8%	33.3%
Neutra	**Enfadada**	0%	0%	0%	2.4%	95.2%	2.4%

Tabla 86 Resultados de identificación de emociones generadas mediante el método de síntesis por copia.

Como se puede ver, el enfado en frío no parece estar prosódicamente marcado, aunque su prosodia es diferente a la neutra, pero no parece transmitir emoción o ser identificable prosódicamente. Salvo en el caso de la sorpresa, el resto de las emociones simuladas no parecen ser claramente reconocibles por su ritmo y su entonación exclusivamente (al menos si la tomamos de un original). Así podemos clasificar el enfado amenazante de nuestro actor como segmentalmente reconocible, la sorpresa como prosódicamente reconocible, mientras que su alegría y su tristeza son mixtas (la tristeza tiene una componente prosódica predominante, mientras que en la alegría predomina lo segmental).

Empleando el análisis y modelado prosódico antes descrito, se creó un módulo prosódico automático para síntesis con emociones. Esta vez empleamos los párrafos para calcular la prosodia automática y aplicarla (re-síntesis) a los difonemas de una frase corta.

Sintetizada como...	Identificada como...					
	neutra	alegre	triste	sorprendida	enfadada	otra
neutra	72.9%	0	15.7%	0	0	11.4%
alegre	12.9%	65.7%	4.3%	7.1%	1.4%	8.6%
triste	8.6%	0	84.3%	0	0	8.6%
sorprendida	1.4%	27.1%	1.4%	52.9%	0	17.1%
enfadada	0	0	0	1.4%	95.7%	2.9%

Tabla 87 Resultados de identificación de emociones generadas mediante re-síntesis automática de prosodia.

Las diferencias entre este experimento final y el primero son significativas (empleando un test *chi-square* con 4 grados de libertad y $p>0.95$), fundamentalmente debido a los bajos resultados de la sorpresa. Analizadas una a una con el test de la T de Student, las diferencias no son significativas ($p<0.05$). La prosodia de los párrafos de sorpresa parece no estar tan marcada como en las frases cortas (los párrafos de la sorpresa no habían sido evaluados previamente).

5.6.1 Conclusiones sobre síntesis de voz con emociones mediante síntesis por copia

En este importante experimento hemos determinado la naturaleza segmental o prosódica de las emociones simuladas: en una evaluación pionera confirmada por posteriores experimentos de otros grupos de investigación, hemos mezclado los difonemas y la prosodia de diferentes emociones para concluir la naturaleza segmental del enfado amenazante del actor de nuestra base de datos y la naturaleza prosódica en el caso de la sorpresa; para la alegría y la tristeza se ha revelado como de naturaleza mixta, en parte segmental en parte prosódica.

Capítulo 6 Conclusiones y líneas futuras

A lo largo de esta Tesis hemos abordado diversas soluciones para 3 subproblemas importantes a la hora de poder dotar de mayor naturalidad y variedad a la conversión texto habla: el procesado lingüístico automático orientado a prosodia, el modelado de la curva de F0 en un dominio y la síntesis de voz personalizada con emociones.

6.1 Conclusiones

6.1.1 Procesado lingüístico automático

§ Se han probado tres técnicas de desambiguación contextual gramatical:

- una basada en reglas manuales (cuyo coste de desarrollo no se ve compensado con una tasa adecuadamente elevada);
- otra basada en reglas inferidas automáticamente (que supere la dificultad de generar manualmente las reglas). En esta técnica de aprendizaje de reglas los resultados han sido excelentes, aunque bastante dependientes del dominio de entrenamiento (pudiéndose reducir la tasa desde un 99% a menos de un 96%), debido al sobre entrenamiento y el aprendizaje de reglas no generales;
- otra basada en modelado estocástico, constatándose el superior comportamiento de esta última técnica al realizar ensayos fuera de dominio. Al emplear la técnica estocástica con diccionarios no adaptados a un dominio concreto sin probabilidades, se ha alcanzado una cobertura del 99,89% en textos de un dominio distinto al dominio de entrenamiento, comparable a los mejores sistemas en castellano (aunque casi todos sobre un corpus distinto) y que (a pesar de no tener probabilidades) supera significativamente en precisión los resultados de un sistema léxico basado en probabilidades, aunque a costa de una mayor ambigüedad media (>96% en el primer candidato). En la desambiguación, resulta también significativa la mejora debida al tratamiento de las locuciones, dado que los modelos probabilísticas basados en categorías no son capaces de modelar bien contextos amplios y con pocos ejemplos.

§ Se ha adaptado un sistema de análisis por medio de gramáticas de contexto libre, desarrollando y evaluando con éxito una gramática robusta de dominio general en dos niveles, uno sintagmático y otro relacional. Cabe destacar el empleo de reglas de corte para reducir el número de análisis posibles (con sólo un 0,35% de imprecisión), la aplicación de reglas de concordancia como filtrado posterior al análisis y el uso de un criterio muy simple de número mínimo de segmentos para elegir el mejor análisis (sin necesidad de información probabilística). En el primer nivel sintagmático los resultados han sido excelentes (cobertura y precisión superiores al 96%, comparables a los mejores resultados en castellano, aunque sobre distinto corpus), a pesar de que haya un 1% de errores debidos al etiquetado previo. En el segundo nivel relacional los resultados son prometedores (cobertura y precisión superiores al 87%).

§ En el nivel léxico, se ha experimentado con diccionarios adaptados al dominio, con diccionarios generales y con diccionarios extranjeros, destacando la aportación de los dos primeros tipos. También se ha constatado la necesidad de incluir reglas robustas para etiquetar palabras fuera de vocabulario, experimentándose con mejoras significativas el empleo de reglas de experto basadas en las terminaciones de las palabras. En este sentido se ha ensayado el empleo de varios conjuntos de reglas manuales de experto procedentes de otros proyectos (completadas con algunas nuevas reglas), aunque la inclusión de los diccionarios ha obligado a filtrarlas y adaptarlas, incrementando su precisión de un 77,5% a un 98,88% (aplicada a un 24,8% de las palabras desconocidas). Se ha incorporado un

conjugador verbal de gran cobertura basado en un paradigma sencillo pero efectivo; a pesar de la sobregeneración de este módulo, no se han producido importantes incrementos en el número de etiquetas por palabra.

§ Trabajando en el nivel de palabra, se han estudiado los distintos tipos de palabras no estándar y la manera de detectar su presencia en un texto, de manera que sea posible procesar no sólo un corpus convenientemente preparado, sino textos obtenidos directamente del dominio sin supervisión. Se ha creado y evaluado un normalizador de texto basado en diccionarios genéricos y diccionarios especializados y en reglas de experto empotradas. La precisión global alcanzada fue del 98,41%, mientras que la precisión sobre las palabras no estándar fue siempre superior al 85% sobre un corpus de evaluación de textos periodísticos, destacando el 96% en nombres propios simples. Para la detección de palabras y nombres propios extranjeros se ha probado un sencillo método basado en las reglas de silabicación del castellano, cuya precisión supera el 99,5%, aunque cubre pocos casos.

6.1.2 Modelado de F0 en dominio restringido

§ Se ha estudiado el modelado mediante perceptrones multicapa, destacando la significativa importancia que adquieren la información sobre "la frase portadora" y el "signo de puntuación final del grupo fónico". El parámetro "número de frase portadora" introduce mejoras significativas; a pesar de que en las condiciones de grabación se intentó aislar el elemento variable de su frase portadora por medio de pausas obligatorias. El parámetro más importante, el "signo de puntuación final del grupo fónico", es muy relevante porque permite distinguir elementos variables con cadencias y elementos variables que generalmente presentan anticadencias en las grabaciones.

§ Se ha constatado la importancia de codificar la información enventanada sobre el acento de cada sílaba, así como su situación inicial o final en el grupo fónico. Se han ensayado varias codificaciones alternativas para la misma información sin conseguir superar la tasa. El tamaño óptimo de la ventana es dependiente de la tarea, aunque tiene relación con el tamaño de los grupos fónicos y los datos disponibles. Los resultados obtenidos son en general coherentes con los resultados de (J.A. Vallejo 1998), aunque ahora ensayados en un dominio variado que va desde el habla aislada hasta los sintagmas nominales largos.

§ Se ha desarrollado y evaluado un nuevo método de diseño de bases de datos: por medio de un nuevo algoritmo voraz moderado por medio de subobjetivos parciales, se ha conseguido resumir una gran base de datos con una precisión superior al 95 %, de acuerdo con amplio espectro de vectores prosódico y segmentales.

§ Es igualmente importante estudiar cómo agrupar las frases en subdominios, (realizar una correcta agrupación de las grabaciones de acuerdo con su prosodia, proponiendo un modelado individual para algunas grabaciones), aunque las diferencias encontradas no han sido significativas.

§ Parámetros secundarios a la hora de modelar han resultado ser el tamaño del grupo fónico en sílabas o en palabras, la pertenencia de la sílaba a una palabra función o su situación en posición final de palabra. Apenas aportan mejoras; y si las aportan nunca es significativamente ni en todos los subdominios.

§ Hemos empleado una estrategia de experimentación no exhaustiva, que al ser comparada con la búsqueda exhaustiva del óptimo, ha mostrado su validez.

6.1.3 Análisis y síntesis de voz con emociones

§ Se han realizado experimentos para determinar la naturaleza segmental o prosódica de las emociones simuladas: en una evaluación pionera confirmada por posteriores experimentos de otros grupos de investigación, hemos mezclado los difonemas y la prosodia de diferentes emociones para concluir la naturaleza segmental del enfado amenazante del actor de nuestra base de datos y la naturaleza prosódica en el caso de la sorpresa; para la alegría y la tristeza se ha revelado como de naturaleza mixta, en parte segmental en parte prosódica.

§ Se ha creado un sistema completo de conversión texto a voz en castellano, con una nueva voz configurable con emociones: para ello se ha empleado síntesis basada en formantes en castellano, con capacidad de personalización evaluada con usuarios. Los parámetros de personalización han sido elegidos de manera que permitan implementar las emociones como un caso particular de personalización dinámica. Por lo que hemos podido ver, los resultados globales resultan prometedores, puesto que los humanos parecen adaptarse con rapidez a la voz sintética con emociones, y el periodo de adaptación podría resultar breve y por lo tanto satisfactorio, especialmente en el caso de la tristeza (siendo peor para el enfado). Es cierto que hubo problemas de inteligibilidad en algunos contextos fonéticos, pero los resultados son perfectamente aceptables, incluso si la voz evaluada no resulta totalmente natural. De acuerdo con los resultados de evaluación, la voz sintética con emociones desarrollada en el proyecto VAESS es comparable al estado de la cuestión en otros idiomas a nivel mundial. Aunque en los planteamientos iniciales del proyecto VAESS se consideró que eran independientes los módulos prosódico y segmental del sintetizador, la conclusión de nuestro trabajo dista mucho de ser esta: debemos señalar que las trayectorias de los formantes pueden provocar, al incrementarse la velocidad de elocución, ruidos de naturaleza pseudo-oclusiva, que es necesario limar uno a uno, modificando la reglas segmentales.

§ Creación de la primera base de datos de habla emotiva simulada en castellano: Está orientada a síntesis prosódica y al análisis de la prosodia en párrafos y frase cortas mediante técnicas paramétricas, y dio lugar a un modelado diferencial de cada emoción respecto a la voz neutra y su evaluación en experimentos de copy-synthesis.

6.2 Líneas futuras

6.2.1 Procesado lingüístico automático

6.2.1.1 Categorización automática

Una primera línea de trabajo sería ensayar la combinación de clasificadores que, al emplear diferentes parámetros de predicción o diferentes ventanas aplicadas a estos parámetros, resulten así complementarios. Se han realizado experimentos preliminares de combinación de métodos de aprendizaje de reglas como complemento al método estocástico, pero es necesario experimentar con nuevos patrones.

Otra posible línea buscaría incorporar probabilidades al modelado léxico generalista para mejorar sus prestaciones dentro de un dominio concreto, y experimentaría con técnicas de adaptación a dominio.

Finalmente, dentro de la categorización automática se podrían aplicar medidas de confianza para la determinar del número óptimo de categorías por palabra que proporcionar al módulo sintáctico con un determinado nivel de confianza.

6.2.1.2 Análisis sintáctico

En cuanto a análisis sintáctico, se debería mejorar el modelado relacional y estudiar su aplicación al pausado en sistemas de conversión texto habla. También se propone entrenar o desarrollar técnicas menos costosas de análisis sintáctico orientado a *chunks*, basadas en clasificadores o en autómatas finitos probabilísticas.

6.2.1.3 Análisis semántico o conceptual

En este campo (no tratado en esta Tesis), se podrían aplicar tanto las técnicas de aprendizaje automático de reglas (para etiquetado semántico), como el análisis sintagmático (con un clasificador que determine el concepto al que se asocia cada sintagma simple detectado).

6.2.2 Modelado de F0 en dominio restringido

6.2.2.1 Nuevo método voraz para diseño de bases de datos

En el diseño de bases de datos podría modificarse el algoritmo para mejorar la cobertura de los parámetros menos frecuentes de los vectores de distribución deseados, potenciando que sean los primeros en ser alcanzados.

También habría que ampliar el algoritmo para tener en cuenta restricciones que tengan que ver con la frecuencia de aparición de estructuras sintácticas.

6.2.2.2 Modelado de F0

Para la misma locutora se debería proceder al modelado de una base de datos prosódica más general y comparar los resultados.

Otra posible línea futura consistiría en ensayar técnicas mixtas basadas en perceptrones multicapa y en técnicas de modelado paramétrico (como el modelo TILT, el modelo de Fujisaki o el modelado con curvas de Bézier), combinando ambos tipos de técnicas para aprovechar las ventajas de cada una. De nuevo la combinación de clasificadores

6.2.3 Análisis y síntesis de voz con emociones

6.2.3.1 Síntesis de voz configurable y con emociones

Sería muy interesante disponer de una demo o un tutorial que expliquen los parámetros y las consecuencias de su modificación, no siempre fácilmente perceptibles, o disponer de diversas configuraciones previas estándar que reflejasen los requerimientos más usuales de los usuarios, que a partir de una voz base más cercana sus intereses, pudiese alcanzar resultados satisfactorios en menor tiempo.

De cara al empleo de síntesis por concatenación personalizable y con emociones, es necesario estudiar las diversas técnicas de conversión de voces.

Otra línea sería la de investigar el empleo de síntesis por selección de unidades aplicada a la síntesis emotiva.

6.2.3.2 Base de datos de habla emotiva en castellano

Por último se podría experimentar con nuevas base de datos, especialmente una que sea multilocutor a fin de poder generalizar las conclusiones del análisis.

Referencias

1. *S. Abney 1996* “Chunk stylebook” working draft http://www.sfs.nphil.uni-tuebinger.de/ abney/ Papers.html#96i.

2. *S. Abney 1997* “Part of speech tagging and partial parsing” en “Corpus-based Models in Language and Speech Processing” Ed. Kluwer.

3. *L. Aguilar, J.M. Fernández, J.M. Garrido, J. Llisterri, A. Macarrón, L. Monzón y M.A. Rodríguez 1994* “Diseño de pruebas de evaluación de habla sintetizada en español y su aplicación a un sistema de conversión de texto a habla” en Actas de SEPLN.

4. *E.Alarcos - RAE 1995* “Gramática de la Lengua Española” Ed. Espasa-Calpe.

5. *J. Alcina & J.M. Blecua 1975* “Gramática española” Ed. Ariel.

6. *J. Atserias, J. Carmona, I, Castellón, S. Cervell, M. Civit, L. Márquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé & J: Turmo 1998* “Morphosyntactic analysis and parsing of unrestricted Spanish text” en

Proceedings of LREC.

7. *J. Bachenko & E. Fitzpatrick 1990* "A computational grammar of discourse-neutral prosodic phrasing in English" en Computational Linguistics 16: 155-170.

8. *A. Batliner, R. Kompe, A. Kiessling, M. Mast, H. Niemann & E. Nöth 1998* "M = Syntax + Prosody: A syntactic-prosodic labeling scheme for large spontaneous speech databases" en Speech Communication 25: 193-222.

9. *M.E. Beckman & G.M. Ayers 1994* "Guidelines for ToBI Labelling" http://www.ling.ohio-state.edu/research/phonetics/ E_ToBI/.

10. *H. Berthelsen & B. Megyesi 2000* "Ensemble of classifiers for Noise Detection on PoS tagged corpora" en Lecture notes in Computer Science. Ed. Springer-Verlag.

11. *O. Boëffard & F. Emerard 1997* "Application-dependent prosodic models for text-to-speech synthesis and automatic design of learning database corpus using genetic algorithm" en Eurospeech Proceedings IV: 2507-2510.

12. *A. Botinis, B. Granstrom & B. Moebius 2001* "Developments and paradigms in intonation research" en Speech Communication 33: 263-296.

13. *T. Brants 1999* "Tagging and Parsing with cascaded Markov Models" PhD Thesis, University of Saarlandes.

14. *T. Brants 2000* "TnT A statistical Part-of-Speech tagger" en Proceedings of ANLP.

15. *G. Brassard & P. Bratley 1996* "Algorithmics: Theory and Practice" Ed. Prentice Hall.

16. *E. Brill & J. Wu 1998* "Classifier combination for improved lexical disambiguation" en Proceedings of COLING-ACL.

17. *E. Brill 1993* "A Corpus-Based Approach to Language Learning " Doctoral thesis, University of Pennsylvania.

18. *E. Brill 1995* "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging" en Computational Linguistics.

19. *J. Cahn 1989* "Generating expression in synthesized speech" Informe técnico del Technology Media Lab del MIT.

20. *N. Campbell 1992* "Syllable-based segmental durations" en "Talking machines" Ed. North Holland.

21. *N. Campbell 2000* "Databases of emotional speech" en Proceedings of the ISCA Workshop: Speech and Emotion.

22. J. Carmona, S. Cervell, L. Márquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé & J: *Turmo 1998* "An environment for morphosyntactic processing of unrestricted Spanish text" en Proceedings of LREC.

23. *D. Casacuberta 2000* "¿Qué es una emoción?" Ed. Crítica.

24. *D. Casacuberta, L. Aguilar & R. Marín 1997* "Propause: a syntactico-prosodic system designed to assign pauses" en Proceedings de Eurospeech 97, vol I pp. 203-206.

25. *A. Casas -Guijarro 1997* "Sistema telefónico multilínea con reconocimiento de voz y acceso a una base de datos remota" PFC ETSI Telecomunicación UPM, dirigido por J. M. Montero.

26. *J. Cassell 2000* "Nudge Nudge Wink Wink: Elements of face-to-face conversation for embodied conversational agents" en "Embodied conversational agents" Ed. MIT Press.

27. *J. Castillo 2000* "Modelado de la fuente de excitación glotal mediante análisis-síntesis LPC en un entorno gráfico de ventanas" PFC ETSI Telecomunicación UPM, dirigido por J. Gutiérrez-Arriola.

28. *J.P. Chanod & P. Tapanainen 1995* "Tagging French – Comparing a statistical and a constraint-based method" en Proceedings of ACL-EACL.

29. *J.P. Chanod & P. Tapanainen 1996* "A robust finite-state parser for French" en Proceedings of ESSLLI Robust Parsing Workshop.

30. *N. Chomsky & M Halle 1968* "The sound pattern of English" Ed. Harper & Row.

31. *N. Chomsky 1959* "On certain formal properties of grammars" en Information and Control.

32. *N. Chomsky 1994* "The Minimalist Program" Ed. MIT Press.

33. *K. Church 1988* "A stochastic parts program and noun phrase parser for unrestricted text" en Proceedings of the 2nd Conference on Applied Natural Language Processing ACL.

34. *M. Civit & I. Castellón 1998* "Gramesp: una gramática de corpus para el español" en la Revista de AESLA.

35. *M. Civit, I. Castellón & M. A. Martí 2001* "Creación, etiquetación y desambiguación de un corpus de referencia del español" en la Revista SEPLN 27: 21-28.

36. *J. Colás 1999* "Estrategias De Incorporación De Conocimiento Sintáctico Y Semántico
En Sistemas De Comprensión De Habla Continua En Español" Tesis doctoral de la ETSI Telecomunicación UPM.

37. *R. Córdoba, J.A. Vallejo, J.M. Montero, J. Gutiérrez-Arriola, M.A. López & J.M. Pardo 1999* "Automatic Modeling of Phoneme Duration in a Spanish Text-to-Speech System Using Neural Networks" en Eurospeech Proceedings IV: 1619-1622.

38. *R. Córdoba, J.M. Montero, J. Gutierrez-Arriola, J. Pardo 2001* "Duration Modeling In A Restricted-Domain Female-Voice Synthesis In Spanish Using Neural Networks" en ICASSP Proceedings.

39. *R. Cornelius 2000* "Theoretical approaches to emotion" en Proceedings oh the ISCA Workshop: Speech and Emotion.

40. *J. Cussens 1997* "POS tagging using Progol" en Inductive Logic Programming: Proceedings of the 7th International Workshop (ILP-97). LNAI 1297, pages 93-108.

41. *Daedalus X.* "Tecnología de Daedalus en Ingeniería Lingüística" http://www.daedalus.es.

42. *W. Daelemans & J. Zavrel 1996* "MBT: a Memory based Part of Speech Tagger Generator" en Proceedings of the Workshop on Very Large Corpora: 14-27.

43. *W. Daelemans, A. van den Bosch & J. Zavrel 1999* "Forgetting exceptions is harmful in language learning" en Machine learning 34: 11-41.

44. *L. Dehaspe & L. De Raedt 1997* "Mining association rules in multiple relations" en Proceedings of the 7th International Workshop on Inductive logic Programming en Lecture Notes in artificial Intelligence 1297: 125-132 Ed. Springer-Verlag.

45. *V. Demonte 1991* "Teoría sintáctica: de las estructuras a la rección" Ed. Síntesis.

46. *E. Dermatas & G. Kokkinakis 1995* "Automatic stochastic tagging of natural language texts" en Computational Linguistics 21: 137-163.

47. *E. Douglas -Cowie, R. Cowie & M. Shröeder 2000* "A new emotion database: considerations, sources and scope" en Proceedings of the ISCA Workshop: Speech and Emotion.

48. *I. Engberg, A. Hansen et al 1997* "Design, recording and verification of a Danish Emotional Speech Database" en Eurospeech Proceedings.

49. *E.V. Enríquez, A. Romero y Juan M. Montero 1996* "VAESS – propuesta de frases para la base de datos" Informe Interno del GTH TR/GTH-DIE-ETSIT-UPM/1-96.

50. *E.V. Enríquez, C. Casado & A. Santos 1989* "La percepción del acento en español" en Lingüística española actual 11:241-269.

51. *D. Escudero-Mancebo, V. Cardeñoso & A. Bonafonte 2002* "Corpus Based Extraction of Quantitative Prosodic Parameters of Stress Groups in Spanish" en Proceedings of ICASSP 2002.

52. *G. Erbach 1994* "Bottom-up Earley deduction" en Proceedings of COLING-94.

53. *G. Fant, A. Kruckenbberg & A. Botinis 2001* "Prominence correlates. A study of Swedish" en Proceedings of Eurospeech.

54. *D. Farwell 2001* "Spanish language processing at CRL" en SLPLT Proceedings.

55. *H. Fujisaki, S. Ohno et al 1994* "Analysis and synthesis of accent and intonation in standard Spanish" en

Proceedings of the ICSLP 355-358.

56. *J.M. Garrido 1991* "Modelización de patrones melódicos del español para la síntesis y reconocimiento de habla" Ed. UAB.

57. *J.M. Garrido 1996* "Modelado de la entonación en español para aplicaciones de conversión texto habla" Tesis doctoral UAB.

58. *J.P. Gee & F. Grosjean 1983* "Performance structures: a psycholinguistic and linguistic appraisal" en Cognitive Psychology 15: 411-458.

59. *D. Gibbon & U. Gut 2001* "Measuring speech rhythm" en Proceedings of Eurospeech.

60. *F. Giménez de los Galanes 1995* "Síntesis de voz de alta calidad en castellano" Tesis doctoral ETSI Telecomunicación UPM.

61. *J. Goldsmith 2001* "Unsupervised Learning of the Morphology of a Natural Language" en Computational Linguistics 27(2) pp 153-198.

62. *D. Goleman 1995* "Emotional intelligence" Ed. Bantan Books.

63. *J.M. Goñi 1998* "Arquitectura para representación del conocimiento léxico en sistemas de procesamiento de lenguaje natural" Tesis Doctoral ETSI Telecomunicación UPM.

64. *J.M. Goñi, J.C. González y A. Moreno 1997* "A lexical platform for engineering Spanish processing tools" en Natural Language Engineering Journal 3: 317-345.

65. *M. González del Campo 2000* "Síntesis de voz por selección de unidades. Aplicación a síntesis de dominio restringido" PFC ETSI Telecomunicación UPM, dirigido por J. M. Montero.

66. *Á.L. González, J. M. Goñi y J. C. González 1995* "Un Analizador Morfológico para el Castellano basado en Chart" en Actas de la VI Conferencia de la Asociación Española para la Inteligencia Artificial.

67. *J.C. González, J.M. Goñi & A. Nieto 1995* "ARIES: a ready for use platform for engineering Spanish processing tools" en Language Engineering Convention.

68. *J. Graña 2000* "Técnicas de análisis sintáctico robusto para la etiquetación de lenguaje natural" Tesis doctoral en la Universidad de la Coruña.

69. *C. Gussenhoven & T. Rietveld 1998* "On the speaker-dependence of the perceived prominence of F0 peaks" en Journal of Phonetics 26: 371-380.

70. *J. Gutiérrez -Arriola 2001* "A new multi-speaker formant synthesizer that applies voice conversion techniques" en Eurospeech Proceedings.

71. *J. Gutiérrrez-Arriola 2001* "New rule-based data driven strategy to incorporate Fujisaki's F0 model to a text-to-speech system in castillian Spanish" En ICASSP Proceedings, pp. 821-824.

72. *J. Hansen & S. Bou-Ghazade 1997* "Getting started with SUSAS: a Speech Under Simulated and Actual Stress database" en Eurospeech Proceedings.

73. *M.S. Harris & N. Umeda 1987* "Difference limens for fundamental frequency contours in sentences" en JASA.

74. *J. Healey & R. Picard 1998* "Digital Processing of Affective Signals" en ICASSP Proceedings.

75. *M. Hepple 2000* "Independence and commitment: assumptions for rapid training and execution of rule-based POS taggers" en Proceedings of the ACL Meeting.

76. *J. Heras 2002* "Utilización de agentes animados para interfaces avanzadas de ayuda" PFC ETSI Telecomunicación UPM, dirigido por J. M. Montero.

77. *R. Herman 1996* "Phonetic markers of global discourse structure in English" en Journal of Phonetics 28: 466-493.

78. *M.L. Herranz y J.M. Brucart 1987* "La sintaxis" Ed. Cátedra.

79. *K. Hirose, M. Eto, N. Minematsu & A. Sakurai 2001* "Corpus-based synthesis of fundamental frequency

contours based on a generation process model" en Proceedings of Eurospeech.

80. *J. Hirschberg & P. Prieto 1996* "Training intonational phrasing rules automatically for English and Spanish text-to-speech" en Speech Communication 18: 281-290.

81. *J. Hopcroft & J. Ullman 1979* "Introduction to automata theory" Ed. Addison Wesley.

82. *I. Iriondo, R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J.M. Blanco, D. Bernadas J.M. Oliver D.l Tena & L. Longhi 2001* "Validation Of An Acoustical Modeling Of Emotional Expression In Spanish Using Speech Synthesis Techniques" en Proceedings of the ISCA Workshop: Speech and Emotion.

83. *A. Jiménez Pozo 1999* "Adaptación y mejora de un sistema de pre-procesamiento y Etiquetado morfosintáctico gramatical" PFC ETSI Telecomunicación UPM, dirigido por J. M. Montero.

84. *H. Jiménez & G. Morales 2001* "Noun phrases identification from a tagged text" http://delta.sc.investav.mx/gmorales/pln/frano3.

85. *D. Johnson & S. Lappin 1997* "A critique of the Minimalist Program" en Linguistics and Philosophy 20: 273-333.

86. *I. Karlsson. 1994* "Controlling voice quality of synthetic speech" en Proceedings of International Conference on Spoken Language Processing, pp. 1439-1442.

87. *M. Kienast & W. Sendlmeier 2000* "Acoustical analysis of spectral and temporal changes in emotional speech" en Proceedings of the ISCA Workshop: Speech and Emotion.

88. *D. Klatt 1973* "Discrimination of fundamental frequency contours in synthetic speech: Implications for models of speech perception" en JASA 53: 8-16.

89. *D. Klatt 1987* "Review of text-to-speech conversion for English" en JASA:82: 737-793.

90. *K. Koskenniemi 1983* "Two-level Morphology: A General Computational Model for Word-Form Recognition and Production" PhD Thesis, University of Helsinki.

91. *J. Laver 1980* "The phonetic description of voice quality" Ed. Cambridge University Press.

92. *G. Leech, M. Weisser, A. Wilson y M. Grice 1998* "Survey and guidelines for the representation and annotation of dialogue" Informe del proyecto EAGLES.

93. *N. Lindberg & M. Eineborg 1998* "Learning constraint Grammar style disambiguation rules using Inductive Logic Programming" en COLING-ACL Proceedings.

94. *J. Llisterri, R. Marín, C. de la Mota & A. Ríos 1995* "Factors affecting F0 peak displacement in Spanish" en Eurospeech Proceedings.

95. *E. López 1993* "Estudio de técnicas de procesado lingüístico y acústico para sistemas de conversión texto-voz en español basados en concatenación de unidades" Tesis doctoral ETSI Telecomunicación UPM.

96. *C. Lyon & B. Dickerson 1995* "A fast partial parse of natural language sentences using a connectionist method" en 7th Conference of the European Chapter of the ACL.

97. *Q. Ma. M. Murata, M. Utiyama, K. Uchimoto & H. Isahara 1999* "Part-Of-Speech tagging with mixed approaches of neural networks and transformation rules" en First Workshop on Natural Language Processing and Neural Networks (NLPNN'99), pp. 53-57.

98. *M.J. Makashay, C.W. Wightman, A.K. Syrdal, & A. Conkie 2000* "Perceptual evaluation of automatic segmentation in text-to-speech synthesis," en ICSLP 2000, vol. II, pp. 431-434.

99. *L. Márquez & L. Padró 1997* "A flexible POS tagger using an automatically acquired language model" en Proceedings of Joint. EACL/ACL 97.

100. *L. Márquez 1999* "POS tagging: a machine learning approach based on decision trees" Tesis doctoral en la UPC.

101. *A. Martín & J.M. Goñi 1995* "Una propuesta y un etiquetador de codificación morfosintáctica para corpus de referencia en lengua española" en Actas del XIII Congreso Nacional de la Asociación Española de Lingüística

Aplicada.

102. *G. Martínez -Salas 1998* "Adaptación de un modelo de duraciones y entonación para sintetizar habla con emociones" PFC ETSI Telecomunicación UPM, dirigido por J. M. Montero.

103. *F. Martínez-Sánchez, J.M. Montero, J. de la Cerra 2002* "Sesgos cognitivos en el reconocimiento de expresiones emocionales de voz sintética en la alexitimia" en Psicothema 14(2), 344-349.

104. *B. Megyesi 2001* "Comparing data-driven learning algorithms for PoS tagging of Swedish" en Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2001) pp. 151-158.

105. *B. Merialdo 1994* "Tagging English text with a probabilistic model" en Computational Linguistics 20: 155-171.

106. *P. Mertens & C. Allessandro 1995* "Pitch contour stylization using a tonal perception model" en ICPhS Proceedings.

107. *A. Mikheev 1997* "Automatic rule induction for unknown word guessing" en Computational Linguistics 23 (3), 405-424.

108. *H. Mixdorf 1998* "Intonation patterns of German Model-based Quantitative Analysis and Synthesis of F0 contours" PhD Thesis, Technischen Universität Dresden.

109. *A. Molina, F. Pla, L. Moreno & N. Prieto 1999* "APOLN: a partial parser of unrestricted text" en 5th Conference on Computational Lexicography and Text Research COMPLEX-99. Proc. pp 101-108.

110. *J.M. Montero 1992* "Desarrollo de un entorno para el análisis sintáctico de una lengua natural. Aplicación al castellano" PFC ETSI Telecomunicación UPM.

111. *J.M. Montero 1998* "Informe sobre la nueva versión del categorizador gramatical basado en diccionarios y reglas" Informe interno TR/GTH-DIE-ETSIT-UPM/2-98.

112. *J.M. Montero, J. Gutiérrez-Arriola, R. Córdoba, E. Enríquez, J.M. Pardo 2002* "The role of pitch and tempo in Spanish emotional speech: towards concatenative synthesis" en "Improvements in speech synthesis" Ed. Wiley & Sons.

113. *J.M. Montero y J. Gutiérrez Arriola 2000* "Proyecto Mejora de calidad de voz femenina: base de datos para prosodia" Informe interno TR/GTH-DIE-ETSIT-UPM/3-00.

114. *J.M.. Montero, R. Córdoba, J.A. Vallejo, J. Gutiérrez-Arriola, E. Enríquez, J.M. Pardo 2000* "Restricted-Domain Female-Voice Synthesis in Spanish: from Database Design to ANN Prosodic Modelling" en ICSLP Proceedings.

115. *J M. Montero 2000* "Análisis sintagmático por medio de gramáticas robustas de contexto libre" Informe interno TR/GTH-DIE-ETSIT-UPM/2-00.

116. *J.M. Montero, J. Gutiérrez-Arriola, J. Colás, E. Enríquez & J.M. Pardo,1999* "Analysis and Modeling of Emotional Speech in Spanish" en Proceedings of the XIVth International Congress of Phonetic Science II: 957-960.

117. *J.M. Montero, J. Gutierrez-Arriola, S. Palazuelos, E. Enriquez, S. Aguilera, J.M. Pardo 1998* "Emotional Speech Synthesis: From Speech Database to TTS" en Proceedings of ICSLP.

118. *J.M.. Montero, L.F. D'Haro, R. Córdoba, J.A. Vallejo, J. Gutiérrez-Arriola, J.M. Pardo 2003a* "ANN F0 Modeling for Female-Voice Synthesis in Spanish: restricted and non-restricted domains" en ICPhS Proceedings.

119. *J.M. Montero, M.M. Duque 2003b* "ANESTTE: a writer's assistant for a specific purpose language" en Proceedings of Corpus Linguistics.

120. *A. Moreno & J.M. Goñi 1995* "A morphological processor for Spanish implemented in Prolog" en Proceedings of the Joint Conference on Declarative Programming pp. 321-331.

121. *P.J. Moreno 1989* "Improving Naturalness in a Text to Speech System with a new Fundamental Frequency algorithm" en Eurospeech Proceedings pp. 360-363.

122. *Y. Morlec, G. Bailly & V. Aubergé 1997* "Synthesizing attitudes with global rhythmic and intonation contours" en Eurospeech Proceedings.

123. *M. Municio, G. Rojo, F. Sánchez-León & O. Pinillos 2000* "Language resources development at the Spanish Royal Academy" en Proceedings of LREC.

124. *I. Murray & J. Arnott 1993* "Towards the simulation of emotion on synthetic speech: A review of the literature on human vocal emotion" en JASA 93: 1097-1108.

125. *I. Murray & J. Arnott 1995* "Implementation and testing of a system for producing emotion-by-rule in synthetic speech" en Speech Communication 16: 369-390.

126. *I. Murray 1989* "Simulating emotion in synthetic speech" PhD Thesis, Universidad de Dundee.

127. *A. Nogueira, A. Moreno, A. Bonafonte & J. Mariño 2001* "Speech emotion recognition using HMM" en Proceedings of Eurospeech.

128. *S. Ohmo & H. Fujisaki 2001* "Quantitative analysis of the effects of emphasis upon prosodic features of speech" en Proceedings of Eurospeech.

129. *J.C. Olabe 1983* "Sistema para la conversión de un texto ortográfico a hablado en tiempo real" Tesis doctoral ETSI Telecomunicación UPM.

130. *The Onomastica Consortium 1995* "The ONOMASTICA interlanguage pronunciation lexicon" en Eurospeech Proceedings I: 829-832.

131. *M. Ostendorf & K. Ross 1997* "A multi-level model for recognition of intonation labels" en "Computing Prosody" Ed. Springer.

132. *L. Padró 1997* "A hybrid environment for syntax-semantic tagging" Tesis doctoral en la UPC.

133. *S. Palazuelos, S. Aguilera, J.M. Montero & J.M. Pardo* "Report on the evaluation of the emotional voice for Spanish" Informe del GTH dentro del proyecto VAESS.

134. *S. Palazuelos -Cagigas 1994* "Incorporación de mejoras ergonómicas y mecanismos predictivos a un editor orientado a personas discapacitadas" PFC ETSI Telecomunicación UPM, dirigido por J. M. Montero.

135. *S. Palazuelos-Cagigas 2001* "Aportación a la predicción de palabras en castellano y su integración en sistemas de ayuda a personas con discapacidad física" Tesis Doctoral ETSI Telecomunicación UPM.

136. *D. Palmer, J. Burger & M. Ostendorf 1999* "Information extraction from broadcast news speech data" en Proceedings of the DARPA Broadcast News Workshop.

137. *J.M. Pardo et al 1987* "Improving Text to Speech Conversión in Spanish. Linguistic analysis and prosody" en Eurospeech Proceedings.

138. J.M. Pardo, F. Giménez de los Galanes, J.A. Vallejo, M.A. Berrojo, J.M. Montero, E. Enríquez & A. *Romero 1995* "Spanish text-to-speech: from prosody to acoustics" en Proceedings of the International Congress on Acoustics, III: 133-136.

139. *J. Pastor, J.Colás, R. San-Segundo, J.M.Pardo 1998* "An Asymetric Stochastic Language Model Based on Multi-Tagged Words" en Proceedings of 5th International Conference on Spoken Language Processing, ICSLP98.

140. *R. Picard 1997* "Affective computing" Ed. MIT Press.

141. *J. Pierrehumbert 1979* "The perception of fundamental frequency declination" en JASA 66, pp. 363-368.

142. *J. Pierrehumbert 1987* "The Phonology and Phonetics of English Intonation" Ed. Indiana University Linguistic Club.

143. *F. Pla & N. Prieto 1998* "Using Grammatical Inference Methods for automatic Part-Of-Speech tagging" en Proceedings of LREC.

144. *F. Pla, A. Molina & N. Prieto 2001* "Evaluación de un etiquetador morfosintáctico basado en bigramas especializados para el castellano" en Revista de SEPLN 27: 215-221.

145. *D. Polanco 2000* "Evaluación y mejora de un sistema automático de análisis sintagmático" PFC ETSI Telecomunicación UPM, dirigido por J. M. Montero.

146. *T. Portele & B. Heuft 1997* "Towards a prominence-based synthesis system" en Speech Communication 21: 61-

72.

147. *P. Prieto, C. Shih & H. Nibert 1996* "Pitch Downtrend in Spanish" en Journal of Phonetics 24: 445-473.

148. *S. Quazza & H. van der Heuvel 2000* "The use of lexica in text-to-speech systems" en "Lexicon development for speech and language processing" Ed. Kluwer.

149. *A. Quilis 1981* "Fonética acústica de la lengua española" Ed. Gredos.

150. *RAE 1973* "Esbozo de una nueva gramática de la Lengua Española" Ed. Espasa-Calpe.

151. *L. Ramshaw & M. Marcus 1995* "Text chunking using transformation-based learning" en Proceedings of the Third Workshop on Very Large Corpora.

152. *A. Ratnaparkhi 1998* "Maximum Entropy models for Natural Language ambiguity resolution" PhD Thesis, University of Pennsylvania..

153. *S. Rodríguez & J. Carretero 1996* "A Formal Approach to Spanish Morphology: the COES Tools" en Proceedings of XII SEPLN 118-126.

154. *M.A. Rodríguez, J.G. Escalada, A. Macarrón & L. Monzón 1993* "AMIGO: un conversor texto-voz para español" en Boletín de la SEPLN 13: 389-400.

155. *G. Rojo et al* "Base de datos sintácticos del español actual" http://www.usc.es/.

156. *K. Samuel 1998* "Lazy Transformations-Based Learning" en Proceedings ACL.

157. *C. Samuelsson & A. Voutilainen 1997* "Comparing a linguistic and a stochastic tagger" en Proceedings of ACL-EACL.

158. *R. San Segundo, J.M. Montero, R. Córdoba, J.M. Gutiérrez-Arriola 2000* "Stress Assignment in Spanish Proper Names" en Proceedings of ICSLP.

159. *F. Sánchez León 1998* "Anotación Lingüística: CREA" en Jornadas de Presentación CREA-CORDE 31-34.

160. F. Sánchez-León, J. Porta, J.L. Sancho, A. Nieto, A. Ballester, A. Fernández, J. Gómez, L. Gómez, E. *Raigal & R. Ruiz 1999* "La anotación de los corpus CREA y CORDE" en Procesamiento del Lenguaje Natural 25: 175-182.

161. *F. Sánchez-León & A. Nieto 1997* "Retargeting a tagger" en "Corpus annotation: linguistic information from computer text-corpora" Ed. Longman.

162. *F. Sánchez-León 1995* "CRATER Final report" http://www.lllf.uam.es/ docsen/final_report/.

163. F. Sánchez-León, J. Porta, J.L. Sancho, A. Nieto, A. Ballester, A. Fernández, J. Gómez, L. Gómez, E. *Raigal & R. Ruiz 1999* "La anotación de los corpus CREA y CORDE" en Procesamiento de Lenguaje Natural 25: 175-182.

164. *F. Sánchez-León 1997* "Análisis morfosintáctico y desambiguación en castellano" Tesis Doctoral de la Universidad Autónoma de Madrid.

165. *J. Sánchez 2000* " Modelado de prosodia en dominio restringido mediante redes neuronales" PFC ETSI Telecomunicación UPM, dirigido por J. M. Montero.

166. *O. Santana, J. Pérez, Z. Hernández, F. Carreras, G. Rodríguez, L .Losada & J. Duque 2001* "Desarrollos del Grupo de Estructuras de Datos y Lingüística Computacional de la Universidad de Las Palmas de Gran Canaria (GEDLC)" en Actas de SLPLT.

167. *J. van Santen 1994* "Assignment of segmental duration in text-to-speech synthesis" en Computer Speech & Language 8, pp. 95-128.

168. *K. Scherer 2000* "A Cross-Cultural Investigation Of Emotion Inferences From Voice And Speech: Implications For Speech Technology" en ICSLP Proceedings.

169. *H Schmid 1994* "Part-of-speech tagging with neural networks" en Proceedings of COLING 172-176.

170. *G. Schneider & M. Volk 1998.* "Adding manual constraints and a lexicon look-up to a Brill-Tagger for German" en Proceedings of the ESSLLI-98 Workshop on Recent Advances in Corpus Annotation.

171. *M.S. Scordilis & J.N. Gowdy 1989* "Neural Network-based Generation of Fundamental Frequency Contours",

Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP89), pp. 219-222.

172. *E. Selkirk 1994* "Phonology and Syntax: the relation between sound and structure" Ed. MIT Press.

173. *J. Shen, H. Wang, R. Lyu & L. Lee 1999* "Automatic selection of phonetically distributed sentence sets for speaker adaptation with application to large vocabulary Mandarin speech recognition" en Computer Speech & Language 13: 79-97.

174. *J.M. Sosa 1999* "La entonación del español" Ed. Cátedra.

175. *R. Sproat, A. Black, S. Chen, S. Kumar, M. Ostendorf & R. Richards 1999* "Normalization of Non-Standard Words" Presentación final del proyecto WS99.

176. *R. Sproat, J. Hirschberg & D. Yarowski 1992* "A corpus-based synthesiser" en ICSLP pp. 563-566.

177. *S. Stamatatos & J.M. Montero 1998* "Part-of-Speech Tagger for Spanish Text Based on the Brill`s Tagger" Informe interno del Grupo de Tecnología del Habla, Dpto. Ingeniería Electrónica, ETSI Telecomunicación, UPM.

178. *C. Subirats 1998* "Automatic Extraction of Textual Information in Spanish" en Journal of Theoretical and Experimental Linguistics 1:1-13.

179. *A. Syrdal, J. Hirschberg, J. McGory & M. Beckman 2001* "Automatic ToBI prediction and alignment to speed manual labeling of prosody" en Speech Communication 33: 135-151.

180. *J. t'Hart 1974* "Discriminability of the size of pitch movements in speech" en IPO Annual Progress Report 9:56-63.

181. *P. Tapanainen & A. Voutilainen 1994* "Tagging accurately: don't guess if you know" en Proceedings of the 4th Conference on Applied Natural Language Processing pp. 47-52.

182. *P. Taylor & A.W. Black 1998* "Assigning Phrase Breaks from part-of-speech Sequences" en Computer Speech and Language 12, 99-117.

183. *P. Taylor 1994* "The Rise/Fall/Conection model of intonation" en Speech Communication 15: 169-186.

184. *P. Taylor 2000* "Analysis and synthesis of intonation using the Tilt model" en JASA 107: 1697-1714.

185. *N. Tomás 1948* "Manual de entonación española" Ed. Guadarrama.

186. *D. Torre 2001* "Segmentación y Etiquetado Fonéticos Automáticos. Un Enfoque Basado en Modelos Ocultos de Markov y Refinamiento Posterior de las Fronteras Fonéticas." Tesis ETSI Telecomunicación UPM.

187. *S. Tournemire 1997* "Identification and automatic generation of prosodic contours for a text-to-speech synthesis system in French" en Eurospeech Proceedings 191-194.

188. *C. Traber 1992* "F0 generation with a database of natural F0 patterns and with a neural network", in "Talking machines theories, models and designs", Ed. Elsevier Science Publishers B.V.

189. *Hartmut Traunmüller and Anders Eriksson 1995* "The perceptual evaluation of F0-excursions in speech as evidenced in liveliness estimations" en J. Acoust. Soc. Am. 97: 1905 – 1915.

190. *R.L. Trask 1996* "A Dictionary of Phonetics and Phonology" Ed. Routledge.

191. *J. Trouvain & W. Barry 2000* "The prosody of excitement in horse race commentaries" en Proceedings of the ISCA Workshop: Speech and Emotion

192. *J.A. Vallejo 1998* "Mejora de la frecuencia fundamental en la conversión de texto a voz" Tesis doctoral ETSI Telecomunicación UPM.

193. *A. van Sluiter & V. van Heuven 1996* "Spectral balance as an acoustic correlate of linguistic stress" en JASA 100: 2471-2485.

194. *D. van Kujik & L. Boves 1999* "Acoustic characteristics of lexical stress in continuous telephone speech" en Speech Communication 27: 95-111.

195. *H. van Halteren, J. Zavrel & W. Daelemans 1998* "Improving data-driven word-class tagging by system

combination" en Proceedings of COLING-ACL pp. 491-497.

196. *H. van Halteren, J. Zavrel & W. Daelemans 2001* "Improving accuracy in word class tagging through the combination of machine learning systems" en Computational Linguistics 27: 199-230.

197. *J. van Santen & B. Moebius 1998* "Description of the Bell Labs intonation system" en Proceedings of the ESCA Workshop on Speech Synthesis 293-298.

198. *J. van Santen 1992* "Deriving text-to-speech durations from natural speech" en "Talking machines" Ed. Elsevier.

199. *P. van Santen & A. Buchsbaum 1997a* "Methods for optimal text selection" en Eurospeech Proceedings II: 553-556.

200. *P. van Santen & R. Sproat 1997b* "Methods and tools" en "Multilingual Text-to-Speech Synthesis: The Bell Labs Approach" Ed. Kluwer Academic Publishers.

201. *R. Vargas 1996* "Analizador de estilo para textos de inglés técnico" PFC ETSI Telecomunicación UPM, dirigido por J. M. Montero.

202. *J. Vergne 2000* "Trends in robust parsing" en Coling Proceedings.

203. *J. Veronis, P. Di Cristo, F.Courtois & C. Chaumette 1998* "A stochastic model of intonation for text-to-speech synthesis" en Speech Communication 26: 233-244.

204. *J. Vilares, D. Cabrero & M. A. Alonso 2001* "Generación automática de familias morfológicas mediante morfología derivativa productiva" en Revista SEPLN 27: 181-188.

205. *A. Voutiainen & T. Järvinen 1995* "Specifying a shallow grammatical representation for parsing purposes" en Proceedings of the 7th Conference of the European Chapter of the ACL.

206. *A. Voutilanen 1993* "XPTool, a detector of English noun phrases" en Proceedings of ACL Workshop on Very Large Corpora 48-57.

207. *M. Wang & J. Hirschberg 1992* "Automatic Classification of Intonational Boundaries" en Computer Speech & Language.

208. *Y. Xu 1999* "Effects of tone and focus on the formation and alignment of F0 contours" en Journal of Phonetics 27: 55-105.

209. *J. Zavrel & W. Daelemans 1999* "Recent advances in Memory-based Part-Of-Speech Tagging" en Actas del VI Simposio Internacional de Comunicación Social 590-597.

210. *E. Zwicker & H. Fastl 1990* "Psychoacoustics" Ed. Springer-Verlag.

Apéndices

A.1 Procesado lingüístico automático

A.1.1 Etiquetado del 860

A.1.1.1 Nuevo etiquetado del corpus 860

De manera similar a (*M. Civit, I. Castellón & M. A. Martí* 2001), proponemos corregir las normas de etiquetación

de proyectos anteriores:

§ Los posesivos, demostrativos e interrogativos nunca deberían ser adjetivos sino determinantes (actualmente son incompatibles con los artículos)

§ Los numerales (cardinales y ordinales) e indefinidos deberían ser determinantes o adjetivos: varios coches, varios de los coches, coches varios, los varios coches, coche alguno, los 5 coches, el coche 5, el coche número 5, 5 de los coches, etc.

§ Los numerales deberían poseer una etiqueta única independientemente de que desempeñen labor de núcleo (*los 5*) o de modificador (*los 5 muchachos*), de manera similar a los adjetivos.

§ Los adjetivos "nominalizados" por un artículo o determinante (*los ricos*) deberían ser considerados como adjetivos, ya que admiten distribuciones que no admite un nombre, como por ejemplo un modificador adverbial (*los muy ricos*)

§ Sólo se deberían considerar como nombres propios los siguientes términos: nombres y apellidos de personas, marcas o nombres de productos, nombres geográficos o astronómicos (planetas, pueblos, ciudades, provincias, estados, etc.). Cada una de las palabras del nombre complejo de una entidad (*Congreso de los Diputados*) deberían ser etiquetadas como palabras ordinarias, aunque guardarán un rasgo que los identificará como integrantes de un nombre propio.

§ La palabra *no* debería ser clasificada con una etiqueta especial dado que forma parte de estructuras verbales.

§ Los adverbios de tiempo y de lugar con componente deíctica (*siempre, nunca, hoy, ayer, mañana, aquí, allí*) deberían recibir una etiqueta especial por sus peculiaridades de distribución, que los aproximan a los pronombres (*desde siempre, para hoy*).

§ El rasgo *número* se debe asignar siguiendo criterios morfosintácticos y no semánticos (*vuestro* es singular, no segunda persona del plural; grupo es singular, no plural)

§ Los participios se deberían etiquetar como adjetivos especiales de origen verbal cuando realicen funciones de modificador o núcleo nominal (*los antes mencionados*), aunque reciban complementos preposicionales que parezcan ejemplos de rección o complementación típicas de un verbo (*lo esperado por todos, los así llamados*). De la misma manera los infinitivos serán sustantivos de origen verbal en similares circunstancias: en su caso, presencia de determinante (*el correr por las praderas de aquí para allá*).

§ En este nivel no se debería distinguir entre verbos transitivos e intransitivos, ni cualquier otro tipo relacionado con la rección, dado que se trata de un fenómeno sintáctico.

§ No se deberían distinguir pronombres reflexivos o recíprocos, dado que no son distinguibles sino semánticamente.

§ La palabra *medio* puede ser sustantivo (*un medio*), numeral partitivo (*medio caramelo*) o adverbio (*medio dormido*); *media* no puede ser adverbio

§ Los pre-determinantes deberían incluir las estructuras numeral + de, indefinido + de, cada uno de, (casi) tod/oa/s como en dos de los coches, muchos de ellos, casi todos los presentes,...

§ Los interrogativos pueden ser adverbiales (*dónde, cómo, cómo de grandes, cuándo, por qué*), determinantes (*qué, cuál des*) o pronombres (*cuál, cuáles*)

§ No se debería hacer distinción entre artículos indefinidos y adjetivos indefinidos, aunque sí entre adjetivos calificativos (*es cierto*) e indefinidos (*cierto hombre*)

§ Sólo se deberían reconocen 3 tipos de conjunciones: subordinantes (*porque, debido a que*), y coordinantes, y la conjunción *que.*

§ La palabra *que* puede ser pronombre relativo, conjunción o partícula comparativa.

§ Existen cuantificadores como *un grupo de*, *un par de*...

§ Existen intensificadores: más, menos, mucha defensa, mucho más, bastante menos, algo más, muy, un poco.

A.1.1.2 Formato de las etiquetas del 860

Cada etiqueta se compone de 10 rasgos o *bytes* no completamente ortogonales. El primero es la categoría primaria; el segundo y el tercero son forman la categoría secundaria; el 4º y 5º *bytes* contemplan el tiempo y el modo verbales, aunque para las categorías primarias adverbio y adjetivo contienen el grado; el rasgo 6 es el número o la persona verbal; el séptimo *byte* especifica el tipo de locución a la que pertenece la palabra, si la hubiere; el octavo rasgo es el género; finalmente los 2 últimos *bytes* reflejan que pronombre enclítico contiene la palabra..

<table>
<tr><th>1</th><th colspan="2">2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th></tr>
<tr><td rowspan="2">Categoría primaria</td><td colspan="3" rowspan="2">Categoría secundaria</td><td>Tiempo verbal</td><td>Modo verbal</td><td>Persona verbal</td><td rowspan="2">Tipo de locución</td><td rowspan="2">Género</td><td colspan="2" rowspan="2">Enclíticos</td></tr>
<tr><td colspan="2">Grado adjetival o adverbial</td><td>Número</td></tr>
</table>

A.1.1.3 Categorías primarias y secundarias

A.1.1.3.1 VERBO

1	2,3	4	5	6	7	8
Verbo **V**	Impersonal **16**	Presente **0**	Infinitivo **0**	1ª singular **I**	Locución verbal **9**	Femenino **F**
	Haber **28**	Pasado **4**	Indicativo **1**	2ª singular **U**	**## ..**	Masculino **M**
	Ser **29**	Imperfecto **8**	Imperativo **2**	3ª singular **H**		N 0 ## ..
	Estar **30**	Futuro **A**	Subjuntivo **3**	1ª plural **W**		
	## ..	Gerundio **G**	Condicional **4**	2ª plural **Y**		
		## ..	Participio **6**	3ª plural **T**		
			## ..	**## ..**		

Los posibles valores del campo enclíticos son: *me* (**00**), *te* (**01**), *se*, *le*, *les*, *lo*, *los*, *la*, *las*, *nos*, *os* (**10**), *mele*, *meles*, *mela*, *melas*, *melo*, *melos*, *tele*, *teles*, *tela*, *telas* (**20**), *telo*, *telos*, *sele*, *seles*, *sela*, *selas*, *selo*, *selos*, *seme*, *sete* (3**0**), *senos*, *seos*, *nosle*, *nosles*, *noslo*, *noslos*, *nosla*, *noslas*, *semelo*, *semelos* (**40**), *semela*, *semelas*, *semele*,*semeles*, *setelo/setelos*, *setela/setelas*, *setele/seteles* (**47**). También es posible, como siempre, que este campo adopte un valor no especificado (**..** o **##**).

A.1.1.3.2 Sustantivo

1	2,3	4	5	6	7	8	9,10
Sustantivo **N**	Común **00**	## ..	## ..	Singular **S**	Locución sustantiva **5**	Femenino **F**	N 0 ## ..
	Propio **06**			Plural **N**	**## ..**	Masculino **M**	
	## ..			N 0 ##		N 0 ## ..	

A.1.1.3.3 Adjetivo

1	2,3	4,5	6	7	8	9,10
Adjetivo **A**	Indefinido **07**	Comparativo **02**	Singular **S**	Locución adjetiva **4**	Femenino **F**	N 0 ## ..
	Calificativo **11**	Superlativo **04**	Plural **N**	**## ..**	Masculino **M**	
	Numeral cardinal **12**	## ..	N 0 ##		N 0 ## ..	
	Numeral ordinal **13**					
	Distributivo **21**					
	## ..					

A.1.1.3.4 Adverbio

1	2,3	4,5	6	7	8	9,10
Adverbio **B**	Lugar **00**	Comparativo **02**	N 0 ##	Locución adverbial **8**	N 0 ## ..	N 0 ## ..
	Tiempo **01**	Superlativo **04**		## ..		
	Modo **03**	## ..				
	Relativo **06**					
	Interrogativo **07**					
	Cantidad **08**					
	Negación **09**					
	Cualitativo **21**					
	## ..					

A.1.1.3.5 Pronombre

1	2,3	4,5	6	7	8	9,10
Pronombre **R**	Personal sujeto **00**	## ..	Singular **S**	## ..	Femenino **F**	N 0 ## ..
	Personal objeto tónico **01**		Plural **N**		Masculino **M**	
	Personal átono **02**		N 0 ##		N 0 ## ..	
	Posesivo **05**					
	Demostrativo **11**					
	Indefinido **14**					
	Numeral cardinal **17**					
	Numeral ordinal **19**					
	Relativo **20**					
	Interrogativo **22**					
	## ..					

A.1.1.3.6 Preposición

1	2,3	4,5	6	7	8	9,10
Preposición **P**	Simple **00**	## ..	N 0 ##	Locución prepositiva **6**	N 0 ## ..	N 0 ## ..
	Contracción **03**			## ..		
	## ..					

A.1.1.3.7 Determinante

1	2,3	4,5	6	7	8	9,10
Determinante **D**	Artículo definido **00**	## ..	Singular **S**	Locución cuantificadora **3**	Femenino **F**	N 0 ## ..
	Artículo indefinido **01**		Plural **N**	## ..	Masculino **M**	
	Posesivo **06**		N 0 ##		N 0 ## ..	
	Demostrativo **08**					
	Preartículo **02**					
	Relativo **03**					
	Interrogativo					
	## ..					

A.1.1.3.8 Conjunción

1	2,3	4,5	6	7	8	9,10
Conjunción **C**	Copulativa/ disyuntiva **02**	## ..	Singular **S**	Locución conjuntiva **7**	Femenino **F**	N 0 ## ..
	Distributiva **04**		Plural **N**	## ..	Masculino **M**	
	Que **06**		N 0 ##		N 0 ## ..	
	Final **07**					
	Temporal **08**					
	Causal **09**					

	Consecutiva **10**					
	Condicional/modal **11**					
	Concesiva **13**					
	Adversativa **19**					
	Ilativa **20**					
	## ..					

A.1.1.3.9 Interjección

Sólo posee valor definido el campo 1 (**I**).

A.1.1.3.10Miscelánea

Sólo posee n valor definido el campo primario (**M**) y los secundarios 2 y 3:

Expresión extranjera (**00**), Número romano (**01**), Número no romano (**02**), Abreviatura (**03**), Sigla (**04**), punto (**06**), coma (**07**), cerrar interrogación (**08**), cerrar admiración (**09**), punto y coma (**10**), dos puntos (**11**), guión (**12**), abrir comillas (**13**), cerrar comillas (**14**), abrir comilla (**15**), abrir paréntesis corchetes o llaves (**16**), cerrar paréntesis corchetes o llaves (**17**), tanto por ciento (**18**), dólar (**19**), *ampersand* (**20**), arroba (**21**), sostenido (**22**), asterisco (**23**), más (**24**), igual (**25**), barra (**26**), menor (**27**), mayor (**28**), cerrar comilla (**29**), circunflejo (**30**), puntos suspensivos (**31**), fechas (**32**), horas (**33**), letras o números con guiones (**34**), abrir interrogación (**35**), abrir admiración (**44**), fin de frase (**46**), letra aislada (**50**), número (**55**), número, letras y puntos (**56**), no especificado (**..**).

A.1.1.3.11Ambigua

Aunque en el corpus 860 no existen palabras ambiguas en cuanto a su categoría primaria, el procesado léxico y contextual hace conveniente definir una etiqueta sin desambiguar completamente (**L**), con las siguientes subcategorías:

Posible letra o número romano (**00**), posible letra (**01**), sustantivo o adjetivo (**02**), sustantivo o verbo (**03**), adjetivo o verbo (**04**), sustantivo o adjetivo o verbo (**05**) determinante o pronombre (**06**).

A.1.2 Lista de paradigmas irregulares empleados

El módulo de conjugación empleado incluye los 3 paradigmas regulares del castellano (-ar, -er e –ir) y los siguientes modelos de conjugación irregular:

-acer (*hacer*)
-andar (*andar*)
-asir (*asir)*
-au1ar (*aunar*)
-au2ar (*aullar*)
-avergonzar (*avergonzar*)
-caber (*caber*)
-caer (*caer)*
-car (*comunica*r)
-cer (*carece*r)
-cir (*luci*r)
-cocer (*cocer*)
-dar (*dar*)
-decir (*decir*)
-e1ar (*apretar*)
-e1er (*heder*)
-e1ir (*repetir*)
-e1zar (*comenzar*)
-e2ar (*cerrar*)
-e2er (*perder*)

-e2ir (*mentir*)
-e3ar (*sembrar*)
-e7r (*reír*)
-ebir (*concebir*)
-edir (*pedir*)
-eer (*leer*)
-egar (*negar*)
-egir (*elegir*)
-eguir (*seguir*)
-ehi1ar (*sobrehilar*)
-ei1ar (*descafeinar*)
-eizar (*homogeneizar*)
-ejercer (*ejercer*)
-embaír (*embaír)*
-endir (*rendir*)
-enyir (*reñir*)
-erguir (*erguir)*
-erir (*sugerir*)
-errar (*errar)*
-ervir (*servi*r)
-estar (*estar*)
-estir (*vestir*)
-eu1ir (*reunir*)
-ezar (*empezar*)
-gar (*entregar*)
-ger (*coger*)
-gir (*dirigir*)
-go2ar (*degollar*)
-guar (*averiguar*)
-guir (*distinguir*)
-haber (*haber*)
-henchir (*henchir)*
-iar (*vaciar*)
-ir (*ir)*
-jugar (*jugar)*
-o1ar (*aprobar*)
-o1car (*volcar*)
-o1cer (*torcer*)
-o1er (*mover*)
-o1gar (*colgar*)
-o1ir (*morir*)
-o1zar (*forzar*)
-o2ar (*apostar*)
-o2er (*morder*)
-o2ir (*dormir*)
-o3ar (*encontrar*)
-o7r (*oír*)
-ocar (*trocar*)
-ogar (*rogar*)
-ohi1ar (*amohinar*)
-ohi1ir (*prohibir*)
-oler (*oler)*
-olgar (*olgar)*
-olver (*volver*)
-orcar (*aporcar*)
-pacer (*pacer*)
-placer (*placer)*

-poder (*poder*)
-poner (*poner*)
-querer (*querer*)
-quir (*delinquir*)
-quirir (*adquirir*)
-raer (*raer)*
-rehusar (*rehusar)*
-roer (*roer)*
-saber (*saber*)
-salir (*salir*)
-ser (*ser*)
-tener (*tener*)
-traer (*distraer*)
-u8nir (*unir)*
-u9ir (*argüir*)
-uar (*evaluar*)
-ucir (*conducir*)
-uir (*concluir*)
-valer (*valer)*
-vencer (*vencer)*
-venir (*prevenir*)
-ver (*ver*)
-ver2 (*prever*)
-yacer (*yacer*)
-zar (*analizar*)

A.1.3 Patrones del experimento de aprendizaje de reglas de categorización

Los patrones contextuales empleados son:

- la palabra anterior o siguiente tiene asignada la categoría Z
- dos palabras antes o después se ha asignado la etiqueta Z
- una de las dos últimas o siguientes palabras tiene la etiqueta Z
- una de las tres anteriores o siguientes palabras tiene la etiqueta Z
- la palabra anterior tiene asignada la categoría Z y la palabra siguiente la categoría W
- la palabra anterior o la siguiente está etiquetada como Z y dos palabras antes o después la categoría es W
- la palabra anterior o siguiente es W
- dos palabras antes o después está la palabra W
- una de las 2 últimas o siguientes palabras es W
- la palabra actual es W y la anterior o siguiente es Z
- la palabra actual es W y la anterior o siguiente tienen asignada la categoría Z
- la palabra actual es W
- la palabra anterior o la siguiente es W y la anterior o la siguiente han recibido la etiqueta Z
- la palabra actual es W y la anterior o siguiente es V con categoría asignada T

Los patrones léxicos son:

- § al suprimir las últimas letras de la palabra (hasta 5) da lugar a una palabra del diccionario
- § las primeras letras (hasta 4) de la palabra son X
- § añadir un prefijo o un sufijo da lugar a una palabra del diccionario
- § la palabra anterior o siguiente es W
- § contiene la letra Z

A.1.4 Conjuntos de etiquetas del experimento de aprendizaje de reglas de categorización

Como el número de categorías del 860 es elevada, aplicamos una simplificación inicial que reduzca su número.

Categoría	Patrón 860	Abreviatura
Gerundio de 'Ser'	V29G.....	GER_SER
Participio de 'Ser'	V29.6.....	PAR_SER
Infinitivo de 'Ser'	V29.0.....	INF_SER
Otras formas de 'Ser'	V29.....	VRB_SER
Gerundio de 'Haber'	V28G.....	GER_HAB
Participio de'Haber'	V28.6.....	PAR_HAB
Infinitivo de 'Haber'	V28.0.....	INF_HAB
Otras formas de 'Haber'	V28.....	VRB_HAB
Gerundio de 'Estar'	V..G.....	GER_EST
Participio de 'Estar'	V...6.....	PAR_EST
Infinitivo de'Estar'	V...0.....	INF_EST
Otras formas de 'Estar'	V.....	VRB_EST
Otros verbos	V.....	VRB
Sustantivo	N00.....	NN
Nombre propio	N06.....	PNN
Acrónimo	N10.....	ACRN
Adjetivo posesivo	A06.....	POS_ADJ
Adjetivo demostrativo	A08.....	DEM_ADJ
Adjetivo interrogativo	A09.....	INT_ADJ
Adjetivo numeral	A12.....	NUM
Adjetivo ordinal	A13.....	ORD
Otros adjetivos	A.....	ADJ
Pronombre personal sujeto	R00.....	PER_PRO
Pronombre objeto	R01.....	OBJ_PRO
Pronombre objeto	R02.....	OBJ_PRO
Pronombre demostrativo	R11.....	DEM_PRO
Pronombre relativo	R20.....	REL_PRO
Pronombre posesivo	R05.....	POS_PRO
Pronombre interrogativo	R22.....	INT_PRO
Pronombre numeral	R17.....	NUM
Pronombre ordinal	R19.....	ORD
Otros pronombres	R.....	PRO
Adverbio relativo	B06.....	REL_ADV
Adverbio interrogativo	B07.....	INT_ADV
Otros adverbios	B.....	ADV
Preposición	P00.....	PREP
Preartículo	P03.....	PREP_ART
Articulo determinado	D00.....	DEF_ART
Artículo indeterminado	D01.....	IND_ART
Conjunción	C.....	CON
Abreviatura	M03.....	ABRV

Varios	M00.....	MISC
Varios	M01.....	MISC
Varios	M02.....	MISC
Signo de puntuación	M.....	PUNC

Tras unos primeros experimentos se procedió a usar un conjunto más detallado a fin de mejorar los resultados, añadiéndose las categorías:

Categoría	Patrón 860	Abreviatura
Otros gerundios	V..G.....	GEROUND
Otros participios	V...6.....	PARTICP
Otros infinitivos	V...0.....	INFINIT
Adjetivo indefinido	A07.....	INDEF_ADJ
Adjetivo distributivo	A21.....	ADJ_DISTRIB
Adjetivo calificativo	A11.....	ADJ_CALIF
Pronombre indefinido	R14.....	INDEF_PRO
Adverbio de espacio	B00.....	ADV_LUGAR
Adverbio de tiempo	B01.....	ADV_TIEMPO
Adverbio de modo	B03.....	ADV_MODO
Adverbio de cantidad	B08.....	ADV_CANTIDAD
Adverbio negativo	B09.....	ADV_NEGAC
Adverbio de cantidad	B21.....	ADV_CUANT
Conjunción ‘que’	C06.....	CON_QUE
Abreviatura	M03.....	ABRV
Palabra extranjera	M00.....	FOREIGN
Número	M02.....	NUM

A.1.5 Tablas de resultados de los experimentos sobre etiquetado estocástico

Diccionario	Anchura del haz	Cobertura	Ambigüedad	Confianza		Diccionario	Anchura del haz	Cobertura	Ambigüedad
tgb2	r1	0,9570	1.0236	0,0027		tg	r1	0,9533	1,0012
tgb2	r2	0,9695	1,0757	0,0023		tg	r2	0,9661	1,0428
tgb2	r3	0,9744	1,1050	0,0021		tg	r3	0,9719	1,0677
tgb2	r4	0,9764	1,1298	0,0020		tg	r4	0,9739	1,0809
tgb2	r5	0,9776	1,1445	0,0020		tg	r5	0,9754	1,0908
tgb2	r10	0,9828	1,2276	0,0017		tg	r10	0,9811	1,1476
tgb2	r25	0,9872	1,3869	0,0015		tg	r25	0,9863	1,2658
tgb2	R100	0,9957	2,0252	0,0009		tg	r100	0,9955	1,7775
tgb2	r0	0,9969	5,0052	0,0007		tg	r0	0,9979	3,2608

Tabla 88 Experimentos con el conjunto de etiquetas completo, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando unigramas

Diccionario	Anchura de haz	Cobertura	Ambigüedad	Confianza (99%)		Diccionario	Anchura de haz	Cobertura	Ambigüedad
tgb2	r1	0,9784	1,0239	0,0019		tg	r1	0,9746	1,0011
tgb2	r2	0,9889	1,0683	0,0014		tg	r2	0,9856	1,0352
tgb2	r3	0,9910	1,0865	0,0013		tg	r3	0,9885	1,0493
tgb2	r4	0,9925	1,1059	0,0011		tg	r4	0,9900	1,0572
tgb2	r5	0,9931	1,1154	0,0011		tg	r5	0,9910	1,0631
tgb2	r10	0,9948	1,1624	0,0010		tg	r10	0,9932	1,0842
tgb2	r25	0,9956	1,2343	0,0009		tg	r25	0,9948	1,1164
tgb2	r100	0,9964	1,3910	0,0008		tg	r100	0,9964	1,1972

tgb2	r0	0,9969	3,4269	0,0007		tg	r0	0,9979	1,6673

Tabla 89 Experimentos con el conjunto de etiquetas completo, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando unigramas

Diccionario	Anchura de haz	Cobertura	Ambigüedad	Confianza (99%)		Diccionario	Anchura de haz	Cobertura	Ambigüedad
tgb2	r1	0,9609	1,0231	0,0026		tg	r1	0,9558	1,0012
tgb2	r2	0,9731	1,0746	0,0022		tg	r2	0,9681	1,0425
tgb2	r3	0,9773	1,0990	0,0020		tg	r3	0,9729	1,0637
tgb2	r4	0,9791	1,1243	0,0019		tg	r4	0,9750	1,0765
tgb2	r5	0,9804	1,1397	0,0018		tg	r5	0,9765	1,0865
tgb2	r10	0,9853	1,2204	0,0016		tg	r10	0,9819	1,1422
tgb2	r25	0,9901	1,4080	0,0013		tg	r25	0,9874	1,2897
tgb2	r100	0,9983	1,9540	0,0005		tg	r100	0,9960	1,7285
tgb2	r0	0,9994	4,8900	0,0003		tg	r0	0,9982	3,1741

Tabla 90 Experimentos con el conjunto de etiquetas simplificadas, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando unigramas

Diccionario	Anchura de haz	Cobertura	Ambigüedad	Confianza (99%)		Diccionario	Anchura de haz	Cobertura	Ambigüedad
tgb2	r1	0,9823	1,0234	0,0018		tg	r1	0,9771	1,0011
tgb2	r2	0,9923	1,0665	0,0012		tg	r2	0,9874	1,0343
tgb2	r3	0,9938	1,0829	0,0010		tg	r3	0,9894	1,0449
tgb2	r4	0,9950	1,1008	0,0009		tg	r4	0,9910	1,0531
tgb2	r5	0,9957	1,1094	0,0009		tg	r5	0,9920	1,0593
tgb2	r10	0,9973	1,1550	0,0007		tg	r10	0,9939	1,0795
tgb2	R25	0,9981	1,2264	0,0006		tg	r25	0,9954	1,1111
tgb2	r100	0,9990	1,3795	0,0004		tg	r100	0,9969	1,1891
tgb2	r0	0,9994	3,4093	0,0003		tg	r0	0,9982	1,6594

Tabla 91 Experimentos con el conjunto de etiquetas simplificadas, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando unigramas

Diccionario	anchura del haz	Cobertura	Ambigüedad	Confianza (99%)		Diccionario	anchura del haz	Cobertura	Ambigüedad
tgb2	r1	0,9749	1,0000	0,0021		tg	r1	0,9764	1,0000
tgb2	r2	0,9817	1,0208	0,0018		tg	r2	0,9812	1,0150
tgb2	r3	0,9850	1,0347	0,0016		tg	r3	0,9836	1,0255
tgb2	r4	0,9871	1,0454	0,0015		tg	r4	0,9855	1,0336
tgb2	r5	0,9883	1,0547	0,0014		tg	r5	0,9866	1,0406
tgb2	r10	0,9922	1,0838	0,0012		tg	r10	0,9906	1,0619
tgb2	r25	0,9959	1,1358	0,0008		tg	r25	0,9940	1,0977
tgb2	r100	0,9981	1,2578	0,0006		tg	r100	0,9964	1,1899
tgb2	r0	0,9994	4,8900	0,0003		tg	r0	0,9982	3.,741

Tabla 92 Experimentos con el conjunto de etiquetas simplificadas, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando bigramas

Diccionario	Anchura del haz	Cobertura	Ambigüedad	Confianza (99%)		Diccionario	Anchura del haz	Cobertura	Ambigüedad
tgb2	r1	0,9875	1,0000	0,0015		tg	r1	0,9888	1,0000
tgb2	r2	0,9919	1,0137	0,0012		tg	r2	0,9914	1,0082
tgb2	r3	0,9936	1,0221	0,0011		tg	r3	0,9923	1,0134
tgb2	r4	0,9947	1,0288	0,0010		tg	r4	0,9933	1,0179
tgb2	r5	0,9953	1,0344	0,0009		tg	r5	0,9936	1,0217
tgb2	r10	0,9967	1,0531	0,0008		tg	r10	0,9951	1,0336

tgb2	r25	0,9980	1,0857	0,0006		tg	r25	0,9963	1,0515
tgb2	r100	0,9989	1,1495	0,0004		tg	r100	0,9972	1,0869
tgb2	r0	0,9994	3,4093	0,0003		tg	r0	0,9982	1,6594

Tabla 93 Experimentos con el conjunto de etiquetas simplificadas, con procesado especial de locuciones, sobre un conjunto de evaluación de 38.310 palabras, empleando bigramas

Diccionario	Anchura del haz	Cobertura	Ambigüedad	Confianza (99%)		Diccionario	Anchura del haz	Cobertura	Ambigüedad
tgb2	r1	0,9727	1,0000	0,0022		tg	r1	0,9758	1,0000
tgb2	r2	0,9797	1,0204	0,0019		tg	r2	0,9805	1,0150
tgb2	r3	0,9827	1,0338	0,0017		tg	r3	0,9830	1,0254
tgb2	r4	0,9847	1,0442	0,0016		tg	r4	0,9848	1,0329
tgb2	r5	0,9861	1,0536	0,0016		tg	r5	0,9859	1,0395
tgb2	r0	0,9970	5,0053	0,0007		tg	r0	0,9980	3,2608

Tabla 94 Experimentos con el conjunto de etiquetas completo, sin procesado especial de locuciones, sobre un conjunto de evaluación de 38.130 palabras, empleando bigramas

Diccionario	Anchura del haz	Cobertura	Ambigüedad	Confianza (99%)		Diccionario	Anchura del haz	Cobertura	Ambigüedad
tgb2	r1	0,9851	1,0000	0,0016		tg	r1	0,9879	1,0000
tgb2	r2	0,9894	1,0132	0,0014		tg	r2	0,9904	1,0084
tgb2	r3	0,9911	1,0217	0,0012		tg	r3	0,9915	1,0137
tgb2	r4	0,9921	1,0284	0,0012		tg	r4	0,9923	1,0176
tgb2	r5	0,9928	1,0336	0,0011		tg	r5	0,9928	1,0213
tgb2	r10	0,9942	1,0517	0,0010		tg	r10	0,9942	1,0338
tgb2	r25	0,9957	1,0824	0,0009		tg	r25	0,9957	1,0511
tgb2	r100	0,9964	1,1445	0,0008		tg	r100	0,9967	1,0864
tgb2	r0	0,9969	3,4269	0,0007		tg	r0	0,9979	1,6673

Tabla 95 Experimentos con el conjunto de etiquetas completo, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando bigramas

Diccionario	Anchura del haz	Cobertura	Ambigüedad	Confianza (99%)		Diccionario	Anchura del haz	Cobertura	Ambigüedad
tgb2	r1	0,9887	1,0000	0,0014		tg	r1	0,9899	1,0000
tgb2	r2	0,9920	1,0106	0,0012		tg	r2	0,9918	1,0062
tgb2	r3	0,9932	1,0160	0,0011		tg	r3	0,9928	1,0102
tgb2	r4	0,9937	1,0204	0,0011		tg	r4	0,9932	1,0128
tgb2	r5	0,9944	1,0240	0,0010		tg	r5	0,9937	1,0151
tgb2	r10	0,9956	1,0370	0,0009		tg	r10	0,9945	1,0238
tgb2	r25	0,9970	1,0586	0,0007		tg	r25	0,9956	1,0374
tgb2	r100	0,9985	1,1011	0,0005		tg	r100	0,9968	1,0615
tgb2	r0	0,9994	3,4093	0,0003		tg	r0	0,9982	1,6594

Tabla 96 Experimentos y gráfica de cobertura con el conjunto de etiquetas simplificado, con procesado especial de locuciones, sobre un conjunto de evaluación de 37.530 palabras, empleando trigramas

Diccionario	Anchura del haz	Cobertura	Ambigüedad	Confianza (99%)		Diccionario	Anchura del haz	Cobertura	Ambigüedad
Tgb2	r1	0,9645	1,0000	0,0025		tg	r1	0,9598	1,0000
Tgb2	r2	0,9736	1,0263	0,0021		tg	r2	0,9664	1,0201
Tgb2	r3	0,9789	1,0425	0,0019		tg	r3	0,9702	1,0330
Tgb2	r4	0,9816	1,0541	0,0018		tg	r4	0,9719	1,0419
Tgb2	r5	0,9829	1,0619	0,0017		tg	r5	0,9734	1,0500
Tgb2	r10	0,9871	1,0911	0,0015		tg	r10	0,9772	1,0769
Tgb2	r25	0,9898	1,1334	0,0013		tg	r25	0,9801	1,1208

Tgb2	r100	0,9924	1,2090	0,0012		tg	r100	0,9837	1,2065
Tgb2	r0	0,9989	2,0039	0,0004		tg	r0	0,9885	2,6177

Tabla 97 Experimentos con el conjunto de etiquetas simplificado, con procesado especial de locuciones, sobre un conjunto de evaluación del dominio de discapacidad de 22.518 palabras, empleando bigramas

Diccionario	**Anchura del haz**	**Cobertura**	**Ambigüedad**	**Confianza (99%)**		**Diccionario**	**Anchura del haz**	**Cobertura**	**Ambigüedad**
tgb2	r1	0,9658	1.0001	0,0024		tg	r1	0,9623	1.0000
tgb2	r2	0,9710	1.0177	0,0022		tg	r2	0,9670	1.0142
tgb2	r3	0,9743	1.0290	0,0021		tg	r3	0,9693	1.0245
tgb2	r4	0,9769	1.0378	0,0020		tg	r4	0,9713	1.0324
tgb2	r5	0,9785	1.0453	0,0019		tg	r5	0,9726	1.0386
tgb2	r25	0,9867	1.1041	0,0015		tg	r25	0,9787	1.0931
tgb2	r100	0,9905	1.1630	0,0013		tg	r100	0,9820	1.1581
tgb2	r0	0,9989	2.0039	0,0004		tg	r0	0,9885	2.6177

Tabla 98 Experimentos con el conjunto de etiquetas simplificado, con procesado especial de locuciones, sobre un conjunto de evaluación del dominio de discapacidad de 22.518 palabras, empleando trigramas

A.1.6 Reglas léxicas de preprocesamiento para el análisis sintáctico

Unidades léxicas compuestas:

- NPropio1= N07*
- LAdverb= 8*
- LConj= 7*
- LPrep= 6*
- LSust= 5*

Preposiciones:

- PrepDe= "de"
- ContraccAL= "al"
- ContraccDEL= "del"
- Prep= P00* P15* P*

Sustantivos:

- NPropio= N06*
- PalNumero= "número"
- Sust= N* M18* M00* M04* M03* M55* M56* M58*
- Anyo= N05*
- PalHoras= "horas"
- PalMes= "mes"
- PalDia= "día" "días"
- PalNavidad= "Navidad"
- DiaSemana= "lunes" "martes" "miércoles" "jueves" "viernes" "sábado" "domingo"
- Mes= "Enero" "Febrero" "Marzo" "Abril" "Mayo" "Junio" "Julio" "Agosto" "Septiembre"

"Octubre" "Noviembre" "Diciembre"

Adverbios:

- PalCasi= "casi"
- AdverbMas= "más"
- AdverbMenos= "menos"
- AdverbMuy= "muy"
- AdverbNo= "no"
- AdverbTemp= "ayer" "hoy" "luego" "ahora" "después" "nunca" "siempre" "mañana"
- AdverbEspac= B00*
- Adverb= "si" B01* B00* B*

Pronombres:

- PosesPost= R05*
- PronCuanto= "cuánto"
- PronDonde= "dónde"
- PronQuien= "quién" “quiénes”
- PronSi= "sí"
- PronObjeto= R02*
- ObjetoPron= R01*
- PosesPost= R05*
- PronCual= "cual" R22*
- PronSujeto= R00*

Adjetivos:

- Adjet= "demás" "tal" "tales" "mismo" "misma" "mismos" "mismas" "tan" "otro" A* V..46* M50*
- AdjetCada= "cada"
- AdjCualquier= "cualquier"
- AdjCualquiera= "cualquiera"
- AdjetMismo= "mismo" "misma" "mismos" "mismas"
- PalSolo= "solo" "solos" "sola" "solas"
- Cardinal= A12* M02* R17* M01* L00* M57*
- CardinAnt= A10*
- Ordin= A13* R19*
- Indef_NUM= "mucho" "mucha" "muchas" "muchos" "ningún" "ninguno" "ninguna" "alguno" "alguna" "algunos" "algunas" "otro" "otra" "otros" "otras" "varios" "varias" "poco" "poca" "pocos" "pocas"
- Indef= "todos" "otro" A07* R1* D01*

Determinantes:

- ArticIndet= D01*
- Artic= D* A09*
- PreArtTodos= "todos" "todo" "todas" "toda"
- PalCuya= "cuyo" "cuya" "cuyos" "cuyas"
- Demostr= A08* R11*
- Poses= A06*

Comparaciones:

- PalTanto= "tanto" "tanta" "tantos" "tantas"
- PalComo= "como"
- AdverbMas= "más"
- AdverbMenos= "menos"
- PalQue= C06* R20*

Signos de puntuación:

- Ptoycoma= M07* M11* M12* M10* M24* M26*
- Comillas= M13* M14* M16* M17* M44* M09* M45* M35* M27* M28*

Conjunciones:

- ConjCoord= C02* M20*
- PalQue= C06* R20*
- Conj= "si" C02* C*

Formas verbales:

- InfHaber= V2800*
- InfSer= V2900*
- InfEstar= V3000*
- GerHaber= V28G.*
- GerSer= V29G.*
- GerEstar= V30G.*
- ParHaber= V2846*
- ParSer= V2946*
- ParEstar= V3046*
- Infin= V..00*
- Gerundio= V..G.*
- Participio= V..46*
- VerboEstar= V30*
- VerboSer= V29*
- VerboHaber= V28*

- Verbo= V* V30* V29* V28*

A.1.7 Gramáticas de contexto libre empleadas

A fin de poder comprender las reglas desarrolladas, describiremos brevemente el formalismo gramatical empleado. El formato de las reglas es parecido al formato BNF, constando de una parte izquierda (símbolo no terminal o gramatical) y de una parte derecha (símbolos no terminales o símbolos terminales, también llamados léxico-sintácticos), separados por un signo '='. A fin de facilitar la escritura de reglas, pero sin llegar a permitir una potencia que dificulte su lectura, la cadena de símbolo de la parte derecha admite:

- § paréntesis no anidados (términos opcionales que permiten reunir varias reglas en una sola),
- § signos '+' (adjuntos un símbolo lo hacen recursivo por la derecha),
- § signos '*' (equivalente a una combinación de '+' y paréntesis),
- § barras verticales disyuntivas.

En general, las líneas que comienzan con 'ÇÇ' son ejemplos y las que comienzan con '<<' son comentarios, que en un preprocesamiento pueden ser eliminados (comentarios) o incorporados a un fichero de pruebas (ejemplos).

A.1.8 Gramática de primer nivel

A.1.8.1 Secuencia de segmentos o sintagmas simples

O= SEGMENTO+

```
<< todo enunciado es una secuencia de segmentos con o sin sentido; cada
segmento es un sintagma o un seudosintagma. Los segmentos (aquí sintagmas
simples) se organizaban en niveles jerárquicos en la gramática original, de
tal manera que en el nivel superior (ORACION) estaban las modalidades y el
estilo directo; en el siguiente (ORACION1) los modificadores oracionales; en
el inferior (ORACION2) las proposiciones más simples. En esta gramática
final, por compatibilidad, se conserva el esquema pero no su formulación.
## ideas verdes incoloras duermen furiosamente.
## furiosamente duermen incoloras ideas verdes.
```

INSERTABLE=SAdv

```
<< los sintagmas adverbiales gozan de una especial movilidad.
## furiosamente incoloras duermen las ideas.
```

SEGMENTO=SNSinDETNum| SNSinDETAnyo| SNSinDETMes| SNSinDETDia| SNSinDETSemana| SNSinDET-Indef

SEGMENTO=SN| SNNum| SNDia| SNMes| SNCant| SNConMas-Menos| SNAnyo| SNSemana| SNInf| SN-Indef

SEGMENTO=SP_NOM| SPMes| SPAnyo| SPCant| SPHora| SPConMas-Menos| SP-Indef| SAdj

```
<< A estos segmentos, se podrá llegar en el primer análisis, de una, varias
o todas las formas que conduzcan a él. Por ejemplo, a SNComp (=PalTanto SN
PalComo SN) se podrá llegar siempre y cuando, los dos SN no sean sintagmas
preposicionales nominalizados. si no es así, es totalmente probable que la
frase haya sido cortada por alguna preposición de las que la forman.
## la casa de Pepe en Ibiza , amarilla.
```

SEGMENTO= SAdv | VERBO

```
<< el adverbio NO, como el adverbio SI, tienen gran libertad sintáctica
especialmente en las respuestas. El gerundio puede desempeñar funciones
adverbiales, pero estructuralmente ser más complejo (proposicional: admite
negación...). En este nivel no podemos aclarar más. Hemos quitado (AdverbNo)
Gerundio, ya que estaba incluida en VERBO (ambigüedad) SAdv y VERB
corresponden al SAdv y al núcleo verbal con negación y pronombres objetos
antepuestos.
```

SEGMENTO= Relativo | LocucConj | SConj | SPunt

<< Admiten ir seguidos de proposiciones subordinadas.

A.1.8.2 Nexos

SConj= Conj | ConjCoord

Relativo= ConjQue | PronQuien | PronDonde | PronCual

<< Los pronombres relativos ligan oraciones y antecedentes. En este nivel simplemente se marcan y sirven para segmentar, si no, se pueden analizar de una manera no aislada por reglas precedentes.

la casa donde ellos durmieron.

SPunt= Ptoycoma | Comillas

<< Los signos de puntuación se dividen en binarios (tipo comillas) o unarios (tipo punto y coma).

A.1.8.3 Formas verbales

VERBO= (AdverbNo) PronObjeto* VERBO1 | (AdverbNo) PronObjeto* VERBO2

<< Los pronombres han de ser antepuestos. Los que incluyen preposición ('contigo'...) se tratan aparte.

como.
se los como.
no se los como.

VERBO2= VERBO1 SPInf | VERBO1 SPInfde

<< Con esta construcción, se aceptarán en un mismo sintagma, todas aquellas perífrasis verbales o aquellos verbos que rigen una preposición determinada, como por ejemplo:

trata de llegar.
aspira a aprobar.
dejar de ser.

VERBO1= VerboSimple | VerboCompuesto

<< Formas simples y compuestas

come , ha comido , está agotado, yo como.

VerboSimple= Verbo|Gerundio|Participio|VerboHaber|FormasSer|FormasEstar

FormasSer= VerboSer | InfSer | GerSer

<< Formas simples de cualquier verbo

él come , él anda , él está con un amigo.
ser bueno.
siendo bueno.

FormasEstar= VerboEstar | InfEstar | GerEstar

estar malo , es terrible.
estando lesionado no se puede correr.

VerboCompuesto= VerboCompuesto1 | VerboCompuesto2

<< Se dividen los verbos compuestos en dos categorías; en la primera están los sintagmas verbales compuestos por dos o más verbos pero sin estar entre ellos el verbo "haber":

está siendo juzgado.
está cantando.

<< Los sintagmas verbales compuestos correspondientes a la segunda categoría serán aquellos que tengan dos o más verbos en su composición y además se encuentre entre ellos el verbo "haber":

ha estado siendo juzgado.
ha estado cantando.
ha venido.

VerboCompuesto1= VerboSer ComplSer | VerboEstar ComplEstar

VerboCompuesto1= Verbo Infinitivos | Verbo Participio | Verbo Gerundios

es mirado.
está sentenciado.
está cantando.
está siendo juzgado.

<< Las segundas construcciones se refieren a aquellos verbos que rigen infinitivo, participio o gerundio respectivamente:

intentar ganar.
permanecer abierto.
seguir cantando.
se van intentando haber ganado.

VerboCompuesto2= VerboCompHaber | VerboCompInfHaber | VerboCompGerHaber

VerboCompHaber= VerboHaber ParticCompl (VerboCompInfHaber) | VerboHaber ParticCompl VerboCompGerHaber

VerboCompHaber= VerboHaber ConjQue Infinitivos | VerboHaber PrepDe Infinitivos

<< Construcciones con todas las conjugaciones posibles del verbo "haber"

ha habido , ha sido robado, ha estado descalificado , ha estado ganando , ha estado siendo juzgado , han podido mantener , han intentado acabar ganando , hay que intentar ganar, ha de jugar , ha conducido habiendo bebido.

VerboCompInfHaber= InfHaber ParticCompl

<< Construcciones con el infinitivo del verbo 'haber':

haber habido , haber sido robado ,haber estado descalificado , haber estado ganando , haber estado siendo juzgado , haber podido mantener , haber intentado acabar ganando.

VerboCompGerHaber= GerHaber ParticCompl | GerHaber ConjQue Infinitivos

<< Construcciones con el gerundio del verbo haber:

habiendo robado , habiendo sido robado , habiendo estado descalificado , habiendo estado ganando , habiendo estado siendo juzgado , habiendo podido mantener , habiendo intentado acabar ganando , habiendo que jugar mañana.

ComplSer= (AdverbNo) Participio

ComplEstar= Participio | Gerundio | GerHaber (ConjQue) (Infinitivos) | GerSer (ComplSer)

<< Posibilidades que pueden acompañar a los verbos 'ser' y 'estar'

está acabado.
está cantando.
está habiendo que cambiar.

Infinitivos= InfinComp (Gerundio) | InfinComp Participio | VerboCompInfHaber

Infinitivos= InfSer (ComplSer) | InfEstar (ComplEstar) | InfHaber

<< Estas formas son las que pueden ir detrás de un verbo que rige infinitivo

InfinComp= (AdverbNo) Infin (AdverbNo) Infin*

para poder intentar parar.

Gerundios= GerHaber | GerEstar | GerSer | VerboCompGerHaber | Gerundio (Infinitivos)

<< Gerundios que irán detrás de un verbo que rija gerundio como "seguir"

sigue habiendo que ganar.

ParticCompl= ParHaber | ParSer (ComplSer) | ParEstar (ComplEstar)

ParticCompl= Participio (AdverbNo) (Infinitivos) (AdverbNo) (Gerundio)

ParticCompl= Participio (AdverbNo) (Gerundio) (AdverbNo) (Infinitivos)

<< Estructuras de participio y todas las posibles continuaciones de dicho participio.

come , ha comido , está agotado.
está cantando , fue comido.

A.1.8.4 Nombres propios

PROPIOS= PROPIOS_SIMPLES | PROPIOS_COMPLEJOS

PROPIOS_SIMPLES= NPropio1 NPropio*
<< Nombres propios de personas, calles...
José García Vidal.
María José Martínez López.

PROPIOS_COMPLEJOS= PROPIOS-DEL | PROPIOS-DE_LA
<< Algunos nombres propios y algunos apellidos llevan las partículas "del" o "de la" intercaladas en su estructura
María del Carmen.
José de la Puerta.
José Luis de Villalonga.

PROPIOS-DEL= PROPIOS-SIMPLES ContraccDEL PROPIOS-SIMPLES

PROPIOS-DE-LA= PROPIOS-SIMPLES PrepDe (DETart) PROPIOS-SIMPLES

A.1.8.5 Sintagma nominal

SN=SN0
<< Nivel superior del sintagma nominal. No se sigue la numeración de la notación X-barra, sino su complementaria:
<< X0 -> .. X1 ..
<< X1 -> .. X2 ..
<< Originalmente contemplaba el nivel de coordinación de sintagmas nominales, ahora propio de la gramática de relaciones.
la casa de Pepe y el congreso de Ibiza.

SN0=SN1 | SNSinDET0 | SNAdj
<< Todo sintagma nominal admite un SPde ligado a él directamente y que puede contener otros SPde encadenados. Pero esto, va a ser separado a la hora de segmentar (reglas de corte). Sustituye a la regla original: SN0=SN1 (SPde) | SNSinDET0.
<< En ese nivel SN0 se unifican los distintos tipos de nominalizaciones o determinaciones

SN1= DETart-dem SNSinDET11

SN1= DETpos (SAdj) N (NUM) (PosNucleo)
<< Estructuras típica de un sintagma nominal
la quinta maravillosa mesa cinco verde.
más maravillosas mesas cinco.
cinco grandes árboles verdes más.
<< El Indef al final de la primera regla es para tratar construcciones como:
no existe posibilidad remota alguna.

SN1= DETpos Indef (SAdj) N (PosNucleo) | DETpos (NUM) Indef (SAdj)

SN1= DETpos Indef_NUM (SAdj) N (PosNucleo) | DETpos (NUM) Indef_NUM (SAdj)
<< Los indefinidos dan lugar con frecuencia a estructuras nominales como adjetivos que son.
las muchas maravillosas mesas verdes.

SNSinDET=SNSinDET0

SNSinDET0= SNSinDET1

SNSinDET1= MAS-MENOS (SAdj) N (NUM) (PosNucleo) | (SAdj) N (NUM) (PosNucleo) MAS-MENOS

SNSinDET1=SNSinDET11

<< Diferenciamos entre SN que pueden llevar cualquier determinante, de aquellos que precisan de artículo determinado o demostrativo como nominalizador

la quinta.

fue quinta en la prueba.

<< Los SNSinDET11 admiten como determinante el artículo determinado o un demostrativo. Los SNSinDET1, no necesariamente.

SNSinDET11=AdverbTemp | AdverbEspac

<< estos adverbios temporales o de posición tienen comportamientos sintácticos frecuentemente nominales. No admiten SPde salvo subcategorizaciones.

comer, por allí.

aquí no es mi lugar.

SNSinDET11= (SAdj) N (NUM) (PosNucleo)

PosNucleo= SAdj (Indef) | SAdj Indef_NUM | SAdj AdjCualquiera

PosNucleo= Indef | Indef_NUM | AdjCualquiera

<< Estructuras que pueden ir pospuestas al núcleo en un sintagma: un adjetivo o sintagma adjetival, un número o un indefinido, o la palabra "cualquiera".

no hay casa alguna.

Una casa cualquiera.

SNSinDET11= (AdjCualquier) Indef (SAdj) N (PosNucleo) | Indef NUM (SAdj) | Indef SAdj | ArticIndet

SNSinDET11= (AdjCualquier) Indef_NUM (SAdj) N (PosNucleo) | Indef_NUM NUM (SAdj) | Indef_NUM SAdj

<< Los indefinidos permiten participar en estructuras nominales incluso en ausencia de determinante, dado su carácter adjetivo; a veces la estructura puede resultar de naturaleza elíptica.

cualquier otra cosa.

una.

cualquiera.

algunas verdes.

SNSinDET11= (SAdj) N (PosNucleo) PosesPost

SNSinDET11= Indef (SAdj) N (PosNucleo) PosesPost | Indef_NUM (SAdj) N (PosNucleo) PosesPost

<< El posesivo PosesPost es un SAdj de colocación posterior al núcleo adjunto, cada vez menos empleado.

esta maravillosa mesa tuya.

familiares míos.

algunas maravillosas mesas mías.

estas otras mesas mías son mejores.

SNSinDET11= AdverbNo Infin

<< El origen verbal del Infinitivo permite estas construcciones negativas con facilidad

no correr.

SNSinDET11= (NUM_ANT) (SAdj) N (NUM_ANT) (SAdj) PROPIOS (NUM) | (NUM_ANT) PROPIOS N+

<< Sintagmas nominales en los cuales haya un nombre propio como núcleo:

el tercer festival Festimad.

calle San Aquilino 34.

marca Ferrari

anticuario Ángel Barrero.
el viejo arquitecto municipal Juan Añón.
<< La segunda parte de la regla cubrirá estructuras como las del ejemplo:
London Festival Ballet.
<< donde Festival y Ballet no son nombres propios según el preprocesador

SNSinDET11= (N) PalNumero (NUM)
<< Sintagmas nominales con los que se va a poder cubrir el análisis de frases como:
de Instrucción número 10.
la butaca número 33.

SNAdj= DETart_dem (MAS-MENOS) SAdj | DET AdjetMismo (N) (PosNucleo) | DETpos SAdj
<< Sintagmas nominales que tienen un adjetivo como núcleo.
el bueno ganó.
los más grandes.
el menos grande.
los mismos que ganaron.
la misma persona.
su mismo nombre.
mi potencial.

SNCant= (DET) (Indef_NUM) NUM_ANT (PalSolo) (SAdj) N (NUM) (PosNucleo) (MAS-MENOS)|(DET) Indef NUM_ANT (SAdj) N (NUM) (PosNucleo) (MAS-MENOS)

SNCant= (DET) NUM_ANT Indef_NUM (SAdj) N (NUM) (PosNucleo) (MAS-MENOS)|(DET) NUM_ANT Indef (SAdj) N (NUM) (PosNucleo) (MAS-MENOS)

SNCant= Indef NUM_ANT (SAdj) N (NUM) (PosNucleo) (MAS-MENOS)|Indef_NUM NUM_ANT (SAdj) N (NUM) (PosNucleo) (MAS-MENOS)

SNCant= NUM MAS-MENOS
<< Regla que incluirá en un mismo segmento aquel fragmento de frase en que se haga referencia a una determinada cantidad de algo:
dos de los Estados miembros de la OTAN.
unos 1800 millones de pesetas.
sus 1000 millones de dólares los invirtió.
otros mil millones de personas más.
otros mil millones más de personas.
104 más.
un solo marco presupuestario.
mis otros mil preciados millones guardados más.
los otros 10 ladrones.

SNSinDETNum= (AdjetCada) (SAdj) NUM (SAdj) (NUM) (SAdj)
<< en esta regla, porque hay opciones que se permiten y no son válidas.

SNNum= DET SNSinDETNum | Indef_NUM SNSinDETNum | AdjCualquier SNSinDETNum
<< Sintagmas nominales que tienen un numeral como núcleo. Posibles ejemplos son:
el cinco es mi número preferido.
quiero dos.
cada dos preciosos cincos.

SNSinDET-Indef= Indef | Indef_NUM | AdjCualquiera

SN-Indef= DETart_dem SNSinDET-Indef
<< Sintagma creado para facilitar la elaboración de SNNumAMPL y SPNumAMPL, en el segundo nivel de análisis sintáctico, una vez que los SN y y los SNde estén siempre juntos
una de las muchas oficinas.

cada una de las múltiples posibilidades.
cualquiera de las pocas oficinas.
otras de las muchas oficinas.
alguna de las muchas oficinas.

SNSinDETAnyo= (SAdj) Anyo (SAdj)

SNAnyo= DET SNSinDETAnyo

el pasado 1990.
el 1990 pasado.

SNSinDETMes= (PreArtTodos) (SAdj) Mes (SAdj)

SNMes= (PreArtTodos) DET SNSinDETMes | Indef_NUM SNSinDETMes | Indef SNSinDETMes

SNMes= AdjCualquier SNSinDETMes | AdjetCada SNSinDETMes

todo Agosto.
el gran Agosto.
un Marzo ventoso.

SNSinDETDia= (SAdj) PalDia (SAdj)

SNDia= (PreArtTodos) DET SNSinDETDia | Indef_NUM SNSinDETDia | Indef SNSinDETDia

SNDia= AdjCualquier SNSinDETDia | AdjetCada SNSinDETDia

el gran día.
cualquier día lluvioso.
todos los días.

SNSinDETSemana= (SAdj) DiaSemana (SAdj) (NUM)

SNSemana= (PreArtTodos) DET SNSinDETSemana | Indef_NUM SNSinDETSemana | Indef SNSinDETSemana

SNSemana= AdjCualquier SNSinDETSemana | AdjetCada SNSinDETSemana

el gran viernes.
cualquier sábado lluvioso.
todos los lunes.

SNConMas_Menos= DET MAS-MENOS (SAdj) | MAS-MENOS SAdj | Indef_NUM MAS-MENOS | Indef MAS-MENOS

<< Sintagmas nominales que lleven AdverMas o AdverbMenos, pero acompañados de alguna otra partícula. Se hace así, para quitar ambigüedades ya que AdverbMas y AdverbMenos por sí solos son SAdv.

muchos más.
más de 15 años.
más del 30% de médicos de la mayor parte de los países de América.
el más grande de todos los participantes.

<< Lo mismo para la partícula menos.

menos de 15 años.
menos del 30% de médicos.
el menos grande de todos los participantes.

SNInf= (DET) Infinitivos | (DET) VerboCompInfHaber

el buen comer.

<< sería tratada como SNInf. De otra forma, se trataría como SN

el haber podido llegar es suficiente.
ganar es fundamental.

SN1= DETpos AdverbNo Infin

su no correr es producto del cansancio.

SN1= Artic PronCual
<< Relativo CUAL precedido de artículo
algunos de los cuales.

SN1= DETart_dem ConjQue
<< Relativo QUE precedido de artículo o demostrativo
aquellos que.
cinco que.

SN1= DETart N Demostr
<< Demostrativos pospuestos
la maravillosa mesa aquella.

SN1= (PreArtTodos) PronCuanto (N) | PreArtTodos (N) | PreArtTodos PronSujeto

SN1= PronSujeto (AdjetMismo) | PronSujeto PalSolo
<< Pronombres personales sujeto y construcciones con CUANTO
yo mismo.
como yo.
todos cuantos coches ve.
todo tipo.
todas ellas.
ellos solos.

NOMINALIZABLES= SP_NOM
<< Sintagmas preposicionales nominalizables.

SN1=SN-SP

SN-SP= DETart_dem NOMINALIZABLES | ContraccAL NOMINALIZABLES | ContraccDEL NOMINALIZABLES
<< Nominalización y nominalización por contracción de SP. Ambigüedad: DETdem SP |
todos aquellos de verde.
unos de verde.
la de verde.
ese de verde.
del de Madrid.
el de ayudar.
el de todos los días.
el de más de 30 millones de ECUS.

SN1= (PreArtTodos) Demostr
<< Pronombres demostrativos
quiero eso.
quiero todos esos.

N= Sust | LocucSust | PROPIOS

N= PalHoras | PalMes | DiaSemana | PalDia | PalNavidad | PalNumero
<< El infinitivo se ha quitado después de crear el SNInf y el SPInf

NUM= Cardinal | Ordin
<< Los numerales pueden ir solos con muchas preposiciones
hasta cinco.
hasta el quinto.

NUM-ANT= NUM | CardinAnt
<< Hay numerales que no pueden estar al final de la frase sino que deben ir seguidos de alguna partícula como puede ser un nombre o un adjetivo. Estas partículas, serán: un, treinta y un...

A.1.8.6 Sintagma adverbial

SAdv= (AdverbNo) (AdverbMuy) Adverb | (AdverbNo) MAS_MENOS (Adverb) | AdverbNo | ADV_AMPL

SAdv= Adverb MAS_MENOS | Adverb Adverb+

SAdv= (AdverbNo) PalTanto | (AdverbNo) PalComo

SAdv= (AdverbNo) (AdverbMuy) LocucAdverb | (AdverbNo) MAS_MENOS (LocucAdverb)

```
  << Cambio de nomenclatura: AdverbMuy por Adverb1
  << Sintagmas adverbiales complementando a verbos
## frecuentemente.
## muy frecuentemente.
## más frecuentemente.
## más.
## muy frecuentemente cansado.
## muy cansado.
## cansado muy.
```

ADV_COMPL= SAdv | AdverbMuy

```
  << Sadverbiales complementando a adjetivos.
## muy cansado.
```

ADV_AMPL= PalCasi Adverb | PalCasi PreArtTodos

```
## casi totalmente.
## casi todos.
```

MAS_MENOS= AdverbMas | AdverbMenos

A.1.8.7 Sintagma adjetival

SAdj= (ADV_COMPL) AdjetNucleo+ | AdjCualquier (AdjetNucleo) | AdverbNo AdjetNucleo

SAdj= PosesPost

```
  << Se ha cambiado la nomenclatura: SAdj por SAdj0
  << ambigüedad: AdjetNucleo ADV_COMPL |
  << estructura máxima de un S.Adjetival sin Participio ejerciendo de verbo.
## las muy frecuentemente ricas herederas.
## las muy frecuentemente asustadas herederas.
## está cansado muy frecuentemente.
```

AdjetNucleo=Adjet|Participio|PalCasi|AdjetCada|PalSolo

```
## la mesa es verde.
  << Pueden funcionar como adjetivos, los adjetivos normales y también los
  participios de verbos:
## la mesa rota.
## el árbol caído.
```

A.1.8.8 Estructuras con determinante

DET= DETart_dem | DETpos

```
  << Determinante en sentido amplio.
```

DETdem= (PreArtTodos) Demostr (INSERTABLE)

```
  << Determinante basado en demostrativos.
## todas estas motos
```

DETpos= (PreArtTodos) (Demostr) Poses (INSERTABLE) | PalCuya

```
  << Determinante basado en posesivos.
## todas sus motos
```

DETart= (PreArtTodos) Artic (INSERTABLE)| ArticIndet (INSERTABLE)

```
  << Determinantes basados en artículos
## todas las , asustadas , chicas.
```

unas mesas.

DETart_dem=DETart | DETdem
<< Determinantes basados en artículos o demostrativos

A.1.8.9 Sintagmas preposicionales

PREP=(Prep) Prep (PrepDe) (INSERTABLE) | LocucPrep (INSERTABLE) | (Prep) Prep CONTR-DEL (INSERTABLE)
<< La construcción -Prep Prep PrepDe- se puede dar en la locución:
en contra de.
en contra del.

PREP-DE=PrepDe (Prep) (INSERTABLE)

CONTR-AL=ContraccAL (INSERTABLE)

CONTR-DEL=ContraccDEL (INSERTABLE)

SP=ObjetoPron
CONTIGO, CONMIGO .

SP= PREP SN | (Prep) CONTR-AL SNSinDET | Prep CONTR-DEL SNSinDET
cansado hasta para sólo " el baile ".
cansó hasta al " poli " del barrio.
cansado hasta del no cantar.

SPde= PREP-DE SN | CONTR-DEL SNSinDET
cansado para , de amarillo , " correr ".
saliendo de entre los coches.

SP= PREP SAdj | CONTR-AL SAdj
<< Nominalización
se puso hasta amarillo.
le dio al amarillo.

SPde= PREP-DE SAdj | CONTR-DEL SAdj
<< Nominalización
de amarillo.
del amarillo.

SP= PREP NUM | (Prep) CONTR-AL NUM
<< Nominalización
pasa a quinto.
hasta para cinco.
estaba el quinto pero subió rápidamente a primero.

SPde= PREP-DE (Prep) NUM| CONTR-DEL NUM
<< Nominalización
monedas de a cinco.
llegó de quinto.

SP= PREP Relativo | CONTR-AL Relativo
para que te hable.
al que hable.

SPde= PREP-DE Relativo | CONTR-DEL Relativo
de que te habló.
del que hablé.

SP= PREP PronSi (PalSolo) | PREP PronSi AdjetMismo | CONTR-AL AdjetMismo
para sí.
por sí solo.
a sí misma.

SPde= PREP-DE PronSi (AdjetMismo) | CONTR-DEL AdjetMismo

dar de sí.
de sí mismo.

SP-Indef= PREP SNSinDET-Indef | PREP-DE SNSinDET-Indef | PREP SN-Indef | PREP-DE SN-Indef

<< Creado para la construcción de SPNumAMPL

SPDia= PREP-DE SNDia | PREP SNDia

SPDia= PREP-DE SNSinDETDia | CONTR-DEL SNSinDETDia | PREP SNSinDETDia | CONTR-AL SNSinDETDia

de día.
pan del día.
en todos los días.
hasta el día 12.

SPSemana= PREP-DE SNSemana | PREP SNSemana

SPSemana= PREP-DE SNSinDETSemana | CONTR-DEL SNSinDETSemana | PREP SNSinDETSemana | CONTR-AL SNSinDETSemana

de un sábado.
en todo el lunes.
hasta el martes 12 de Agosto.
al Domingo.

SPMesDe= PREP-DE SNMes | PREP-DE SNSinDETMes | CONTR-DEL SNSinDETMes

de cualquier Noviembre.
de Noviembre.
del caluroso Julio.

SPMes= PREP SNMes | PREP SNSinDETMes | CONTR-AL SNSinDETMes

en un Agosto.
para todos los Agostos.
en Agosto.
al caluroso Agosto.

SPAnyoDe= PREP-DE SNSinDETAnyo | CONTR-DEL SNSinDETAnyo

de 1345.
del futuro 2009.

SPAnyo= PREP SNSinDETAnyo | PREP SNAnyo | CONTR-AL SNSinDETAnyo

en 1987.
en el próximo 2010.
al pasado 1900.

SPNavidad= PREP-DE PalNavidad

de Navidad

SPConMas-Menos= PREP (DET) MAS_MENOS (SAdj) | CONTR-AL MAS-MENOS SAdj

SPConMas-MenosDe= CONTR-DEL MAS-MENOS SAdj | PREP-DE (DET) MAS-MENOS (SAdj)

<< los adverbios 'mas' y 'menos' permiten estructuras nominales cuantificadas también en sintagmas preposicionales no siempre determinados.

en los más de 30 días.
en más de 25 años de vida.
de menos de 30 millones.
en menos del 30% de los casos.

SPHora= PREP SNNum PalHoras

<< Este primer tratamiento de las horas debe ser complementado en el nivel sintático

desde las 17 horas a las 18 horas.

hasta las ansiadas 20 horas.
a dos horas de la finalización del plazo.

SPTemp= PREP AdverbTemp | PREP-DE AdverbTemp | CONTR-DEL AdverbTemp
<< Estos adverbios de tiempo son sintácticamente muy especiales, exhibiendo cierto carácter nominal o pronominal
desde hoy.

SPCant= PREP SNCant | CONTR-AL SNCant

SPCantDe= PREP-DE SNCant | CONTR-DEL SNCant
la deuda ascendió a unos 1900 millones de pesetas.
de unos 1000 millones de deuda se pasó a unos 500.

SPNum= PREP SNNum | PREP-DE SNNum | CONTR-AL SNSinDETNum | CONTR-DEL SNSinDETNum

SPNum= PREP SNSinDETNum | PREP-DE SNSinDETNum
entre el 8 y el 20.
nota media de 8,5.
a dos grandes.

SPInf= PREP SNInf | CONTR-AL SNInf

SPInfde= PREP-DE SNInf | CONTR-DEL SNInf
<< Como siempre, distinguimos la preposición no marcada 'de' o, en este caso, 'del'. La contracción 'al', aunque pudiera unirse a 'a' para formar una categoría especial (complemento indirecto), puede ser tratada como las demás preposiciones en rección.
para ganar.
de haber llegado.
al bajar.

SP_NOM= SP| SPde| SPNum| SPDia| SPMesDe| SPAnyoDe| SPNavidad| SPTemp| SPSemana| SPInf| SPInfde| SPConMas_MenosDe| SPCantDe
<< Sintagmas preposicionales que se pueden nominalizar mediante un artículo o determinante demostrativo.
esto de ayudar a alguien.
el de todos los días.
el de más de 30 millones de euros.

A.1.8.10 Locuciones

LocucConj= LConj+
sin embargo.
a pesar de.

LocucPrep= LPrep+
cerca de.
además de.
así como.
al igual que.

LocucAdverb= LAdverb+
al mismo tiempo.
a la par.
de nuevo.

LocucSust= LSust+
coches patrulla.

A.1.9 Gramática de segundo nivel

O= SEGMENTO+

SEGMENTO= SINTAGMA | SINTAGMA_ALTO_NIVEL

<< Los sintagmas, serán aquellos que se han podido conseguir en el análisis sintáctico. Los sintagmas de alto nivel, serán aquellos que se integran en este nivel de la gramática.

SINTAGMA= SINT_NOMINALES | SINT_PREPOSICIONALES | SINT_OTROS

SINT_NOMINALES= NOMINALES_SINDET | NOMINALES

SINT_PREPOSICIONALES= SINT_PREP_DE | SINT_PREP

NOMINALES_SINDET= SNSinDETNum | SNSinDETAnyo | SNSinDETMes | SNSinDETDia | SNSinDETSemana | SNSinDET_Indef

NOMINALES= SN | SNNum | SNDia | SNMes | SNCant | SNConMas_Menos | SNAyo | SNInf | SN_Indef

SINT_PREP_DE= Spde | SPMesDe | SPAnyoDe | SPCantDe | SPConMas_MenosDe

SINT_PREP= SP | SPNum | SPDia | SPSemana | SPMes | SPAnyo | SPNavidad | SPCant | SPHora | SPTemp | SPInf | SPInfde | SP_Indef | SPConMas_Menos

SINT_OTROS= Sadj | Sadv | VERBO | Relativo | LocucConj | Sconj | SPunt

SINTAGMA_ALTO_NIVEL= SINT_NOMINALES_AN | SINT_PREPOSICIONALES_AN | SINT_OTROS_AN

SINT_NOMINALES_AN= SNSPde | SNNumeracion | SNNumAMPL | SNFecha | SNCant_COMPLEX | SNMas_MenosDe | SNComp | SNCoord | SNComillas

SINT_PREPOSICIONALES_AN= SPSPde | SPNumAMPL | SPFecha | SPCant_COMPLEX| SPMas_MenosDe | SPComp | SPCoord | SPComillasDe | SPComillas

SINT_OTROS_AN= SAdjCoord | SAdjComp | VERBO_Coord | SAdv_AN | SPunt_AN

SPSPde= SP SPde+

A.1.9.1 Cuantificación

SNCant_COMPLEX= SNCant SPde (MAS_MENOS) (SPde) | SNCant SPCant (MAS_MENOS) (SPde)

SNCant_COMPLEX= SNCant SPde MAS_MENOS_DE SINT_NOMINALES | SNCant SPCant MAS_MENOS_DE SINT_NOMINALES

dos de los Estados miembros de la OTAN.
unos 1800 millones de pesetas.
sus 1000 millones de dólares los invirtió.
otros mil millones de personas más.
otros mil millones más de personas.
104 más.
un solo marco presupuestario.
mis otros mil preciados millones guardados más.
los otros 10 ladrones.

SNNumAMPL= SNSinDETNum SPde+ | SNNum SPde+ | SNSinDET_Indef SPde+ | SN_Indef SPde+

una de las muchas oficinas.
cada una de las múltiples posibilidades.
cualquiera de las pocas oficinas.
otras de las muchas oficinas.
alguna de las muchas oficinas.

SPCant_COMPLEX= SPCant SPde (MAS_MENOS) (SPde) | SPCantDe SPde (MAS_MENOS) (SPde)

SPCant_COMPLEX= SPCant SPde MAS_MENOS_DE SINT_NOMINALES | SPCantDe SPde MAS_MENOS_DE SINT_NOMINALES

la deuda ascendió a unos 1900 millones de pesetas.
de unos 1000 millones de deuda se pasó a unos 500.

SPMas_MenosDe= SPConMas_Menos (SPCantDe) SPde+ | SPConMas_MenosDe (SPCantDe) SPde+

SPMas_MenosDe= SPConMas_Menos SPCantDe | SPConMas_MenosDe SPCantDe

en más de 30 millones.
en más de 25 años de vida.
en más del 30% de los casos.
<< Lo mismo para la partícula "menos"
en menos de 30 millones.
en menos del 30% de los casos.

SNMas_MenosDe= SNConMas_Menos (SPCantDe) SPde+ | SNConMas_Menos SPCantDe

SNMas_MenosDe= MAS_MENOS (SPCantDe) SPde+ | MAS_MENOS SPCantDe

SNMas_MenosDe= MAS_MENOS_DE SNCantDe SPde+ | MAS_MENOS_DE SNCantDe

muchos más.
más de 15 años.
más del 30% de médicos de la mayor parte de los países de América.
el más grande de todos los participantes.
<< Lo mismo para la partícula menos.
menos de 15 años.
menos del 30% de médicos.
el menos grande de todos los participantes.

SPNumAMPL= SPNum SPde+ | SP_Indef SPde+
<<SPNumAMPL= PREP SNNumAMPL | PREP_DE SNNumAMPL
en una de las muchas oficinas.
por cada una de las muchas casas.
para esa otra de las muchas novias de tu hermano.

MAS_MENOS= AdverbMas | AdverbMenos

MAS_MENOS_DE= AdverbMasDe | AdverbMenosDe

A.1.9.2 Fechas

SPFecha= SPNum SPMesDe (SPAnyoDe)

SPFecha= SPMes SPAnyoDe | SPMesDe SPAnyoDe

SPFecha= SPDia SPNavidad | SPSemana (SPMesDe) (SPAnyoDe)

decreto de 28 de Julio del 2000.
del 15 de Junio hasta el 10 de Julio.
en Febrero de 1993.
válido hasta Septiembre del 2001.
de Agosto de 1975.
del día de Navidad.
al martes 12 de Agosto de 1990.

SNFecha= SNNum SPMesDe (SPAnyoDe) | SNMes SPAnyoDe

SNFecha= SNDia SPNavidad | SNSemana (SPMesDe) (SPAnyoDe)
<< SNFecha= DET (SAdj) PalMes SPMesDe -> para permitir:
el próximo mes de Enero.
<< Sintagmas nominales que analizarán cualquier forma de dar una fecha.
el pasado sábado.
el próximo 5 de Junio.

```
## el caluroso 11 de Agosto de 1943.
## ese Enero del 1322.
## el gran Agosto de 1975.
## su pasado mes de Enero.
## todos los martes de Agosto de 1992.
## el viernes 13 de Abril de 1999.
```

A.1.9.3 Comparaciones

SPComp= PalTanto SP PalComo SP | PalTanto SPde PalComo SPde

SPComp= PalTanto SPInf PalComo SPInf | PalTanto SPInfde PalComo SPInfde

SPComp= PalTanto SPSPde PalComo SP | PalTanto SPSPde PalComo SPSPde

SPComp= PalTanto SNSPde PalComo SN | PalTanto SNSPde PalComo SNSPde

```
## tanto en Enero como en Febrero.
## tanto de caza como de pesca.
## tanto los derechos de ellos como los nuestros.
```

SAdv_AN= PalTanto | PalComo | AdverbMas | AdverbMenos

SNComp= PalTanto SN PalComo SN | PalTanto SNInf PalComo SNInf

```
## tanto Enero como Febrero.
## tanto las rosas rojas como las rosas blancas.
```

A.1.9.4 Coordinación

SNSPde= SN SPde+

SNNumeracion= SN Coma SN SConj SN

SPCoord= SP SConj SP | SP SConj SN | SPAnyo SConj SNSinDETAnyo | SPInf SConj SNInf | SPInfde SConj SPInfde

```
## una jugada letal y efectiva.
## los grandes y los pequeños.
```

SNCoord= SN SConj SN | SN SConj SAdj | SN SConj SNInf | SNInf SConj SNInf

```
## una jugada letal y efectiva.
## los grandes y los pequeños.
```

SAdjCoord= SAdj SConj SAdj

SAdjComp= SAdj SPde+

VERBO_Coord= VERBO SConj VERBO

A.1.9.5 Comillas

SNComillas= (SN) Comillas SN Comillas | (SN) Comillas SNSPde Comillas

```
## un "comité de asuntos de infraestructura de transporte".
## "Los jóvenes europeos".
```

SPComillasDe= SPde Comillas SN Comillas | SPde Comillas SNSPde Comillas

SPComillas= SP Comillas SN Comillas | SP Comillas SNSPde Comillas

```
   ## de una "cota de desviación".
```

SPunt_AN= Coma | Comillas

A.2 Modelado de F0 en dominio restringido

A.2.1 Frases patrón iniciales de la base de datos de dominio restringido

1) La nacional I tiene, en sentido salida de Madrid, en la provincia de Álava, circulación

interrumpida en **_NOMBRE DE POBLACIÓN_,** entre los puntos kilométricos 15 al 20.

2) La nacional I tiene, en Madrid, el puerto de **_NOMBRE DE PUERTO_,** con cadenas.

3) La nacional I tiene, en Madrid, los puertos de **_NOMBRE DE PUERTO_** y **_NOMBRE DE PUERTO_** cerrados.

4) El tren Estrella, **_NOMBRE DE POBLACIÓN_ - _NOMBRE DE POBLACIÓN_**, sale a las **_HORA_** y llega a las **_HORA_.**

5) La última estación por la que ha pasado el tren Estrella, con origen en Irún y destino a Madrid, es **_NOMBRE DE POBLACIÓN_,** y llegará a su destino con 25 minutos de retraso.

6) De acuerdo, Señor **_APELLIDO_,** ¿Qué operación desea?

7) ¿Qué operación desea Señor **_APELLIDO_?**

8) Señor **_APELLIDO_,** ¿Qué operación desea?

9) El último movimiento contabilizado en su cuenta de ahorro es el siguiente: el uno de enero **_MOVIMIENTO_,** 10.000 pesetas.

10) Los últimos movimientos contabilizados en su cuenta de ahorro son los siguientes: el uno de febrero cargo por **_MOVIMIENTO_** , 5.685 Ptas.; el 13 de febrero abono por **_MOVIMIENTO_,** 6.587 Ptas. y finalmente, el 16 de febrero cargo por **_MOVIMIENTO_,** 16.789 Ptas.

11) La próxima vez puede acceder a esta operación diciendo: **_PALABRA CLAVE_**

12) El estado del cheque de 300.000 pesetas es de **_ESTADO DE CHEQUE_.**

13) Se lo enviamos al número **_NÚMERO DE TELÉFONO_.**

14) Repasemos los datos, quiere hacer un traspaso a su **_TIPO DE CUENTA_,** de la oficina 1.234, de la entidad **_ENTIDAD_** por importe de 25.000 pesetas.

15) Diga los cuatro dígitos de la coordenada **_LETRA_-_NÚMERO_.**

16) ¿A su **_TIPO DE CUENTA_?**

17) Repasemos los datos, quiere suscribir en **_FONDO_** 25.000 pesetas.

18) ¿De su **_FONDO_** número 3457?

19) ¿Quiere conocer el valor de la **_DIVISA_?**

20) Un momento, le pasamos con **_NOMBRE Y APELLIDO_**.

21) ¿Desea información sobre **_TIPO DE INFORMACIÓN_?**

22) La A-483, Carretera de **_POBLACIÓN_,** tiene, en sentido ascendente de kilometraje, circulación intensa.

A.2.2 Frases patrón definitivas de la base de datos de dominio restringido

1) La nacional I tiene, en la provincia de Álava, circulación interrumpida en **<nombre de población>,** entre los puntos kilométricos 15 al 20.

2) La nacional I tiene, en Madrid, el puerto de **<nombre de puerto>**, con cadenas.

3) La nacional I tiene, en Madrid, los puertos de **<nombre de puerto>** y **<nombre de puerto>** cerrados.

4) El tren Estrella, **<nombre de población>** - **<nombre de población>**, sale a las 7 y llega a las 8.

5) La última estación por la que ha pasado el tren Estrella, con origen en Irún y destino a

Madrid, es <**nombre de población**>, y llegará a su destino con 25 minutos de retraso.

6) De acuerdo, Señor <**apellido**>, ¿Qué operación desea?

7) Señor <**apellido**>, ¿Qué operación desea?

8) El último movimiento contabilizado en su cuenta de ahorro es el siguiente: el uno de enero <**movimiento**>, 10.000 pesetas.

9) Los últimos movimientos contabilizados en su cuenta de ahorro son los siguientes: el uno de febrero cargo por <**movimiento**> , 5.685 Ptas.; el 13 de febrero abono por <**movimiento**>, 6.587 Ptas. y finalmente, el 16 de febrero cargo por <**movimiento**>, 16.789 Ptas.

10) La próxima vez puede acceder a esta operación diciendo: <**palabra clave**>.

11) El estado del cheque de 300.000 pesetas es de <**estado de cheque**>.

12) Repasemos los datos, quiere hacer un traspaso a su <**tipo de cuenta**>, de la oficina 1.234, de la entidad <**entidad**> por importe de 25.000 pesetas.

13) ¿A su <**tipo de cuenta**>?

14) Repasemos los datos, quiere suscribir en <**fondo**> 25.000 pesetas.

15) ¿De su <**fondo**> número 3457?

16) ¿Quiere conocer el valor de la <**divisa**>?

17) Un momento, le pasamos con <**nombre y apellido**>.

18) ¿Desea información sobre <**tipo de información**>?

19) La A-483, Carretera de <**nombre de población**>, tiene circulación intensa.

A.2.3 Análisis estadístico del modelado de F0 parámetro a parámetro

Un sencillo análisis estadístico de una de nuestras bases de datos de entrenamiento (proveniente de uno de 10 sub-experimentos) nos puede mostrar aproximadamente qué parámetros son, aisladamente, los que más pueden ayudar al perceptrón a predecir. Si calculamos la salida media cuando un determinado parámetros vale 1 y cuando vale 0, y la diferencia entre ambas resulta significativa estadísticamente, ese parámetro es un candidato a ser un buen predictor, dado que valores diferentes de ese parámetro de entrada, se traducen en salida diferentes (en promedio).

Tomando un ejemplo () donde sólo se muestran los parámetros con diferencias significativas en la media de f0 cuando adoptan el valor 0 y cuando adoptan el valor 1:

§ Podemos observar que para todos los valores del contexto existen diferencias significativas, aunque parece que los contextos posteriores son más influyentes, aunque varios de ellos contemplan pocos casos. El contexto principal abarcaría las sílabas entre la –1 y la +2.

§ Tres de las terminaciones (pausa espontánea-coma-, tonema descendente-punto-, tonema ascendente-punto y coma-) dan lugar a diferencias significativas de la frecuencia fundamental promedio, constatándose que las pausas espontáneas y los tonemas ascendentes suben el tono medio de sus grupos fónicos.

§ Parece existir una cierta diferencia entre los grupos fónicos de hasta 7 sílabas (171 Hz.) y los de más de 7 sílabas (178 Hz.).

§ Los mejores predictores (con un número importante de casos positivos y negativos) parecen ser "el acento de la sílaba en cuestión" y "si la siguiente sílaba es acentuada" (lo cual justificaría que la zona final comience en la sílaba anterior a la última tónica, junto al hecho de que sea importante saber si la anterior es o no es final). El parámetro "si la siguiente sílaba es o no es inicial" es casi equivalente a saber si es acentuada o no, dado que abundan las grabaciones con sólo una tónica, y dicha tónica marca el final de la zona inicial. El parámetro "si dentro de 2 sílabas hay una sílaba inicial" equivale a "si la siguiente sílaba es acentuada", codificando sílabas finales, la primera

acentuada y la anterior a la primera tónica (lo que podríamos llamar iniciales pretónicas).

Parámetros con muy poco poder discriminante (menos de 5 Hz de diferencia en media) son: si hace 3 sílabas había una inicial", "si la anterior es Inicial", "si la actual es Inicial", "si la siguiente es Final".

Los parámetros más desequilibrados (aquellos en los que su valor 0 o su valor 1 apenas se dan, en los que hay una relación 20:1 entre un caso y el otro) pueden ser buenos prediciendo, pero contribuirán poco a la mejora global de resultados. Es el caso de Acent-5 (si hace 5 sílabas había una acentuada), Ini-5, Acent-4, Fin-3, Ini+3 (si 3 sílabas después hay una inicial), Ini+4, Acent+4, Ini+5.

Todos estos datos deben ser valorados teniendo en cuenta el predominio de las palabras aisladas en nuestros experimentos.

Tabla 99 Parámetros de predicción de F0 que presentan diferencias significativas al adoptar valor 1 frente a adoptar valor 0

A.2.4 Análisis de F0 con un modelo paramétrico en dominio restringido

A.2.4.1 Nombres propios en enunciativas

Empleando un modelo de picos y valles (*G. Martínez-Salas* 1998), podemos analizar el comportamiento de los principales puntos de la curva de F0 con respecto a parámetros tales como el número de tónicas, la finalización en oxítona o paroxítona o el signo de puntuación final. Como las poblaciones son pequeñas, emplearemos en nuestro análisis un nivel de confianza del 99%.

Tabla 100
Modelado paramétrico de la curva de F0 para los nombres propios con una sola tónica (modelo de picos y valles)

§ **Oxítona/no oxítonas**: para el guión intermedio, las tónicas presentan diferencias significativas de F0 si el grupo fónico contiene una única tónica, aunque en ambos (una o varias tónicas), se mantiene la tendencia a que la frecuencia fundamental sea alta; para el punto, es el fonema final el que experimenta un descenso significativo de F0 al pasar de oxítona a no oxítona, siendo en ambos casos de valor bajo. En el caso de grupos fónicos con varias tónicas, las diferencias significativas sólo se dan en las tónicas cuando acaba el grupo en punto y coma. Si el tonema es ascendente, la palabra oxítona tiene más F0 en la tónica (por hallarse más cerca del final), produciéndose el efecto contrario si el tonema es descendente.

§ **Signo de puntuación**: todos presentan entre ellos alguna diferencia significativa. Los que más se parecen son el guión y la coma, aunque esta última es poco frecuente y son significativamente distintos en la tónica (si palabra final oxítona). Los que menos similitudes presentan son el punto y el punto y coma, que apenas coinciden en el tono inicial del grupo.

Tabla 101
Modelado paramétrico de la curva de F0 para los nombres propios con varias tónicas (modelo de picos y valles)

§ **Número de tónicas**: sólo hay diferencias significativas en el caso del guión y para la primera tónica y el primer valle. Hay que tener en cuenta que se trata de grupo fónicos siempre cortos.

A.2.4.2 Sintagmas nominales en enunciativas

Empleando de nuevo un modelo de picos y valles (*G. Martínez-Salas* 1998), podemos analizar el comportamiento de los principales puntos de la curva de F0:

§ **Oxítona/paroxítonas**: para el punto y coma, las tónicas presentan diferencias significativas de F0 si el grupo fónico consta de una sola sílaba, aunque se mantiene la tendencia a que la frecuencia fundamental sea alta; para el punto, el número de ejemplos es muy bajo.

§ **Signos de puntuación**: todos presentan diferencias entre ellos (la presencia del guión es anecdótica), con poca similitud, aunque coma y punto y coma presentan final ascendente.

§ **Número de sílabas**: el signo más frecuente cuando hay varias tónicas es la coma (pausa

intermedia), no presentando diferencias significativas cuando hay una o más tónicas.

Tabla 102 Modelado paramétrico de la curva de F0 para los nombres propios con varias tónicas (modelo de picos y valles)

Tabla 103 Modelado paramétrico de la curva de F0 para los sintagmas nominales en enunciativas con una tónica (modelo de picos y valles)

A.3 Análisis y síntesis de habla con emociones

A.3.1 Personalización de voz

A.3.1.1 Evaluación inicial del sintetizador (previo a la personalización)

A.3.1.1.1 Fricativas sordas */f/, /TH/, /s/, /X/*

Estos fonemas eran, en general, los más aceptables del sistema, aunque en la mayoría de los casos se implementaba por medio de aspiración y fricación simultáneamente.

- El fonema

/f/ sonaba más natural cuando venía precedido por una /a/, una /e/ o una /i/, aunque la intensidad resultaba excesiva.

- El fonema

/TH/ se percibía como si fuese una /X/ si la vocal inmediatamente anterior era una /a/. La definición de esta consonante delante de /e/ era más bien deficiente. Si la vocal siguiente es /i/, el sonido asemejaba una mezcla entre una /f/ y una /X/.

- La

/s/ sonaba con una especie de oclusión o transición oclusiva, esta explosión era especialmente evidente cuando la vocal no era posterior (o sea, cuando era /a/, /e/, /i/). Delante de la /u/, la /s/ sonaba como /f/. El carácter aspirado era especialmente antinatural en este fonema.

- La

/X/, sonido extraordinariamente característico del español, resultaba muy deficiente, con demasiada aspiración. En el caso de preceder a una /i/, se notaba algún ruido explosivo espurio, propio de una transición abrupta entre estos dos sonidos de diferente naturaleza. Al llevar detrás una /u/, el sonido que debería ser /X/ parecía más bien una /f/.

A.3.1.1.2 Oclusivas sonoras */b/, /d/, /g/*

El sistema de partida carecía totalmente de este tipo de sonidos, siendo sustituidos por oclusivas sordas.

A.3.1.1.3 Aproximantes */B/, /D/, /G/*

La implementación de estos sonidos intervocálicos era más bien pobre, con mucho ruido de aspiración. La

/B/ era muy fricativa (un sonido que resulta no nativo en un castellano estándar peninsular). La /G/ era ligeramente mejor, pero con un sonido muy fuerte, que se percibía como fricativa /v/ al ir seguida por vocales anteriores (/e/, /i/), y que sonaba como /s/ delante de una /u/. El peor de estos sonidos era la /D/, por su gran aspiración y ganancia sonora.

A.3.1.1.4 Africadas */ch/, /y/*

La africada sorda

/ch/ era bastante aceptable, aunque la fase fricativa era excesivamente fuerte para el uso castellano. La sonora /j/ (que se corresponde con una letra "y" en posición inicial) parecía más bien una vocal /i/ ordinaria.

A.3.1.1.5 Laterales */l/, /L/*

La palatal

/L/ no había sido incluida; sin embargo, la /l/ era bastante buena, aunque posiblemente un poco larga en exceso.

A.3.1.1.6 Nasales */n/, /m/, /ñ/*

La palatal

/ñ/ había sido implementada como una secuencia de /n/ y /y/. La nasalidad del resto era excesiva, aunque en todo caso eran claramente identificables.

A.3.1.1.7 Vibrantes */r/, /R/*

La vibrante simple

/r/ resultaba más fuerte que en las grabaciones naturales de que disponíamos, y no se tenía en cuenta el efecto de los diferentes contextos fonéticos. Por su parte, la /R/ resultaba excesivamente velar, como en una pronunciación ligeramente francesa.

A.3.1.1.8 Grupos consonánticos

Algunas combinaciones como

/b/ + /s/ + /t/ o /n/ + /s/ + /t/ resultaban completamente fallidas. Las demás presentaban problemas generales de las oclusivas (sorda en vez de sonora, etc.).

A.3.1.2 Bases de datos para una voz neutra

Las únicas bases de datos externas que se hallaban disponibles para analizar y modelar voz sintética en castellano eran las EUROM. Sin embargo, tras estudiar su contenido, observamos que no todos los contextos fonéticos estaban presentes, no contenían voz con emociones y necesitaban un etiquetado detallado. Como parte del *background* del GTH, se disponía de una base de datos orientada a prosodia castellana y una base de datos segmental orientada a la extracción de difonemas, ambas emocionalmente neutra y grabada por el mismo locutor. A partir de estos datos y de la voz previamente disponible en el GTH (Tel-Eco), hemos podido definir la nueva voz. El único análisis adicional necesario fue la extracción de formantes, empleando herramientas semiautomáticas adaptadas a la base de datos de difonemas.

A.3.1.3 Herramientas semiautomáticas

De los trabajos previos en síntesis de voz realizados en el GTH, disponíamos de un conjunto de herramientas tales como un marcador de *pitch* (instantes de cierre glotal; indirectamente es un extractor de F0) y un extractor de formantes. Todo ello fue incluido en un entorno de edición de voz que incluye, entre otras, las 2 herramientas anteriores, capaz de trabajar sobre plataforma Windows con interfaz GUI y que admite ficheros de audio de formatos diversos y ficheros de gran tamaño (J. Sánchez 2000) y que es ampliamente configurable.

La otra herramienta fundamental ha sido el editor de parámetros EDP, un programa que nos permite modificar manualmente las trayectorias de los parámetros del sintetizador. En este programa fue necesario adaptar el sintetizador de Klatt extendido con parámetros de fuente (D. Klatt et al 1989) para simular el modelo paralelo (también con parámetros glotales) empleado por Infovox. Finalmente se adaptó la salida del editor para su posterior uso con el verdadero sintetizador de Infovox.

A.3.1.4 Diseño e implementación de una nueva voz

Establecimos 4 objetivos en este trabajo de creación de una nueva voz neutra que sirva de base a los posteriores desarrollos en síntesis de emociones:

§ Un correcto preprocesado lingüístico del texto, que corrija los errores antes señalados y contenga los nuevos fonemas necesarios, así como una reestructuración del juego de rasgos empleados en las definiciones.

§ Un nuevo modelado prosódico basado en el acento castellano de nuestra base de datos prosódica: un modelo más flexible y personalizable de F0 (por picos y valles), y un modelo multiplicativo para las duraciones.

§ Mejorar las reglas segmentales para producir, de una manera correcta y perfectamente distinguible, cada fonema en cada contexto posible, lo cual implicaba una revisión y corrección sistemáticas.

§ Incluir nueva reglas relativas a la fuente glotal novedosa incluida en el nuevo sistema GLOVE (que no sólo permite una mejor producción de voz con emociones, sino que también permite incrementar la naturalidad y calidad general de la voz neutra estándar).

En paralelo con este trabajo, dentro del proyecto VAESS otros equipos participantes procedían a trabajar con voz femenina en otros idiomas (sueco, inglés y danés) y a implementar la versión en tiempo real del sistema.

A.3.1.4.1 Sistema de desarrollo de Rulsys

Además de las herramientas adaptadas propias del GTH, en el curso del trabajo empleamos las herramientas propias de Rulsys. La creación de una nueva voz supone la creación o modificación de hasta 10 ficheros que la van a caracterizar y describir:

§ el fichero de símbolos, rasgos y parámetros por defecto (sin tener en cuenta efectos de contexto),

§ el juego de caracteres que se permitirá en el texto de entrada,

§ el fichero de definición de parámetros (que especifica los nombres, rangos y cuantificaciones de los parámetros de frecuencias y anchos de banda),

§ el fichero de conversión de dígitos en fonemas,

§ el fichero de excepciones, expresiones y palabras función,

§ el fichero de sufijos y raíces (innecesarios para pronunciar correctamente en castellano),

§ el fichero de conversión grafema a fonema,

§ el fichero de símbolos BLISS (necesario para la interfaz con el comunicador orientado a personas con discapacidad),

§ el fichero de reglas fonéticas (segmentales y prosódicas).

Todos estos ficheros son compilados y ejecutados por el entorno de desarrollo Hisys, que permite editar visualmente las trayectorias de los parámetros. En un proceso continuo de prueba y error. Se echó de menos el uso del ratón, un *zoom* y un visualizador de espectrograma como el disponible en EDP, motivo por el cual se prefirió usar la versión adaptada de este.

La labor de editar las reglas se veía dificultada por la ausencia de un editor con realce según sintaxis (*syntax highlighting*), dado que los espacios en blanco y los tabuladores resultan significativos (no son simples separadores).

A.3.1.4.2 Metodología

Centrándonos en el trabajo fonético (más que en la parte de ingeniería de preprocesamiento, por lo demás nada novedosa), podemos decir que tras construir el fichero de definiciones, disponemos de un primer prototipo de la voz, muy primitiva, eso sí. Para ello partimos de la voz Teleco.

Para la creación de reglas contextuales analizamos las grabaciones de logatomos (base de datos para síntesis por concatenación de difonemas), y generamos dos tipos de reglas:

§ **Reglas generales contextuales** que especifican las transiciones entre los principales segmentos (sonoras / sordas, nasales / orales o continuas / discontinuas). La mayoría de las posibles transiciones bruscas se evitan por medio de estas reglas y así proporcionan una marco general para las posteriores, minimizándolas, con el consiguiente ahorro de tiempo y mayor elegancia.

§ **Reglas específicas** que, en caso necesario, definen con claridad cada fonema en cada contexto. Son resultado de las deficiencias observadas durante la evaluación en contextos VCV (vocal-consonante-vocal) y VCCV de las reglas generales, haciendo especial énfasis en la inteligibilidad de cada fonema y en que las transiciones carezcan de ruidos indeseados en un amplio margen de velocidades de elocución.

En este trabajo se empleó, con satisfactorios resultados, un sistema de control de versiones como si se tratase de un desarrollo de SW o HW. El proceso de escribir reglas lingüísticas es una tarea de investigación y programación que emplea un lenguaje orientado a tareas fonéticas en vez de un lenguaje de propósito general. El control de versiones puede permitir la coordinación de varias personas trabajando sobre la misma voz, o el desarrollo paralelo de diferentes aspectos de una misma voz (segmentales, de duraciones o de entonación). También sirve para mantener un registro histórico de cambios, para comparar diferentes estadios de desarrollo de la nueva voz o para recuperar una versión previa que carecía de un defecto incorporado a posteriori.

A.3.1.5 Reglas Prosódicas

A.3.1.5.1 Reglas de duración

Las nuevas reglas son una simple adaptación del modelo multiplicativo empleado en Teleco y calculado a partir de la base de datos de prosodia. Los parámetros que determinan la duración de un fonema en un contexto son: el propio fonema, la velocidad de elocución, el acento, el número de sílabas, la estructura de la sílaba (abierta o cerrada), la posición prepausa y el contexto fonético siguiente. Sin embargo, ha sido necesario modificar algunas reglas porque entraban en conflicto con algunas de las reglas segmentales y provocaban transiciones demasiado bruscas.

A.3.1.5.2 Reglas de entonación

El modelo empleado es simple pero flexible, que incorpora los parámetros que luego permitirán personalizar la voz o implementar las distintas emociones (valor medio de F0, rango de F0, pendiente de la curva entonativa).

Los picos de F0 se localizan, en este modelo, en las vocales acentuadas. El primer pico responde a la fórmula 94 +42*<valor medio de F0>/100. Dado que la curva de los valles no resulta perceptualmente muy importante, a estos se les asigna un valor constante. El resto de los puntos se obtiene por interpolación lineal de picos (teniendo en cuenta la duración en tramas en vez del número de picos, como se hacía en Teleco) o interpolación lineal entre picos y valles. La pendiente por defecto es ligeramente descendente (un hertzio cada 25 tramas de 10 milisegundos).

A.3.1.6 Nuevas reglas Segmentales

Entre las reglas segmentales generales del sistema, las menos dependientes de cada fonema en concreto, podemos reseñar:

§ El sintetizador de Infovox presenta un polo corrector en altas frecuencias (FH), y esta presencia puede entrar en conflicto con el cuarto formante F4. A fin de evitar problemas de resonancia ha sido necesario eliminar dicho formante.

§ Los formantes se propagan a través de las pausas, colocándose así las transiciones bruscas en estos silencios.

§ Cuando nos encontramos dos fonemas sordos continuos, las transiciones de formantes de la rama paralelo, comienzan en el primero.

§ Las transiciones de los formantes serie son rápidas cuando enlazan sonidos sordos y sonoros, y comienzan en el fonema sordo.

§ Las transiciones del cero nasal son más bien lentas y deben comenzar antes de que comience la nasal propiamente dicha, a fin de evitar una carencia de definición en la vocal.

§ Las transiciones entre sonoras continuas son siempre lentas y, por lo tanto, suaves.

§ Las transiciones entre continuas de otro tipo son ligeramente más rápidas y comienzan más tarde.

A.3.1.6.1 Oclusivas sonoras: */b/, /d/, /g/*

La característica barra de sonoridad que precede a la explosión no estaba presente (no había ganancia vocálica previa). No presentaban aspiración, y muy poca fricación. Un polo nasal contribuye a incrementar agradablemente su calidad sonora.

Fue necesario modificar los 2 primeros formantes de

/d/ y /g/.

A.3.1.6.2 Oclusivas sordas: */p/, /t/, /k/*

En castellano su característica explosión (*burst*) es bastante cercana a la vocal siguiente, sin aspiración y con poca fricación. De todas maneras resulta fundamental implementar unas transiciones de los formantes bruscas que simulen el fenómeno de la oclusión. En posición después de pausa se aplican reglas especiales, dada la ausencia de contexto anterior.

En el caso de estar precedidas por fonemas sordos, es necesario definir correctamente las transiciones de los formantes de la rama paralelo.

Los valores formánticos de la

/t/ son dependientes de la vocal posterior.

A.3.1.6.3 Aproximantes: */B/, /D/, /G/, /y/*

Se trata de sonidos fuertemente contextuales, bicontextuales incluso (así el primer formante es más alto si el contexto incluye la vocal

/a/), bastante suaves. Estos fonemas no deberían contener aspiración, aunque un poco de fricación contribuye a producir la suavidad que los caracteriza (ello implica la definición de los valores de 2 formantes en la rama paralelo).

La

/G/ posee ahora nuevos formantes serie, más adecuados. La /y/ tiene una transición más larga y una ganancia sonora suave.

A.3.1.6.4 Fricativas sordas: */s/, /f/, /TH/, /X/*

Así como en otros idiomas estos sonidos pueden realizarse con una considerable componente de aspiración, los sonidos castellanos, excepto la

/X/ se caracterizan por una considerable fricación.

La /s/ presenta unos formantes de la rama paralelo que se encuentran más juntos (el primero es más alto y el segundo más bajo), con nuevos anchos de banda que modifican su distribución de energía espectral. El fonema

/TH/ presenta una inserción de formantes retardada, con anchos de bandas menores.

El fonema

/X/ se caracteriza por copiar los formantes de su contexto vocálico, preferentemente del posterior, en caso de existir. El sistema Rulsys no nos permitía poner el primer formante paralelo por debajo de 800 Hz., así que fue necesario incrementar su ancho de banda cuando el

contexto fuese una vocal /o/ o bien /u/, provocando una minimización del ancho de banda del segundo formante.

A.3.1.6.5 Nasales: */n/, /m/, /ñ/*

La nasalización en castellano es algunas tramas menor en el tiempo, más suave en el caso de la alveolar

/n/ y la bilabial /m/, tanto en ganancia sonora como en ganancia nasal. Fue necesario corregir los valores de los principales formantes de la /m/.

El alófono velar del fonema

/n/ no ha sido implementado por no resultar importante en este nivel del modelado.

El sonido palatal

/ñ/ resulta bastante característico en castellano, con transiciones largas y lentas de sus tres primeros formantes.

A.3.1.6.6 Laterales */l/, /L/, /l1/*

La transición entre este tipo de sonido y las vocales adyacentes resulta todavía más rápida que en resto de los fonemas continuos. Así por ejemplo, fue necesario incrementar la velocidad de transición asociada al segundo formante serie F2.

El alófono

/l1/ comprende los sonidos laterales en grupo consonántico, más suaves que en otros idiomas (menor ganancia sonora) y con valores de F1 y F3 distintos.

A.3.1.6.7 Vibrantes */r/, /R/, /r1/*

Los valores de los formantes de estos sonidos son muy dependientes del contexto, tanto si este es vocálico como si es consonántico, factor que hubo que rediseñar en la nueva voz.

Cuando el fonema vibrante forma parte de un grupo consonántico (

/r1/), sólo debemos tener en cuenta el contexto fonético posterior (el segmento vocálico de estas vibrantes es fundamental para la consecución de un correcto sonido castellano). No se debe dotar al sonido de una alta ganancia sonora, a fin de que el sonido resulte algo más suave. Para su implementación no se necesitan ganancias en la rama paralelo.

La vibrante simple

/r/ es ahora algo más larga, con una mayor velocidad de transición. La /R/ incluye un nuevo contorno de ganancia sonora (temporización) y una mayor velocidad en las transiciones que permita conseguir el efecto de vibración múltiple que la caracteriza.

A.3.1.6.8 Africadas: */ch/, /j/*

La inserción de la fricación se retardó porque era excesiva en algunos contextos, y resultaba poco natural y no contribuía a su correcta identificación. El fonema

/ch/ presentaba una cierta aspiración que se eliminó, incrementándose la ganancia de fricación. Los valores de los formantes paralelos fueron, sin embargo, reducidos. Por lo que respecta a la africada /j/, hubo de ser diseñada e implementada desde cero.

A.3.1.7 Integración y pruebas

En el transcurso de la fase de implementación y pruebas en el sistema de tiempo real no se encontraron problemas dignos de mención, a pesar de que los desarrollos se centraron sobre una versión de RULSYS basada en OVE (fuente glotal estándar) y no se dispuso de GLOVE (fuente de Fant), hasta un estadio muy avanzado del trabajo.

A.3.2 Ejemplo de cuestionario para la evaluación de síntesis de voz

con emociones

Marque la emoción que identifique al escuchar cada frase.

FRASE	Neutra	Alegre	Triste	Sorprendido	Enfadado	No identificada	Otros
1							
2							
3							
...							

A.3.3 Textos de la base de datos SES

A.3.3.1 Párrafos

1) Los participantes en el Congreso marcharon después a El Escorial. Se trasladaron allí en un amplio autobús, en el que un guía iba explicando los monumentos relevantes del recorrido. La visita al monasterio fue comentada por el mismo guía que debía saber mucho sobre El Greco, *en cuyo cuadro* "el martirio de San Mauricio" se extendió ampliamente; no debía ser igual su conocimiento del resto de los cuadros que componían la pinacoteca, sobre los cuales pasó como un rayo, dando lugar a sonrisas cómplices.

2) Sergio era un joven serio y trabajador que vivía cerca de la hospedería del Monasterio de Guadalupe, en las Villuercas, comarca perteneciente a la provincia de Cáceres. Se ganaba la vida vendiendo recuerdos alusivos a la Virgen Morenita, desde llaveros a platos con la imagen grabada en esmalte vidriado. Tenía un problema y era que su tiendecita era de mala construcción y estaba en una parte del pueblo muy empinada, fenómeno por otra parte normal en aquel lugar. Había mucho turismo en la zona. Sergio tuvo la mala suerte de perder su tienda en las últimas inundaciones, pues un corrimiento de tierras se la llevó por delante, con lo cual se le acabó su modo de vida.

3) Pablo estudiaba en la Universidad Politécnica de Madrid y estaba deseando regresar a Medellín; echaba de menos los productos de la matanza y los quesos frescos que hacía su abuela. Ya faltaba poco para las vacaciones; entonces volvería a las orillas del Guadiana, bajo los chopos. Su deseo era tan grande que a veces se le hacían años los pocos días que faltaban.

4) La vida diaria a menudo no es tan fácil, aunque estemos en el final del siglo veinte. Sobre todo cuando los dos en la pareja trabajan. Siempre hay que preguntarse si ya se cambió la ropa, si la puerta tiene el cerrojo o si tengo la llave en el bolsillo. Yo llevo al niño en el coche. Todos los días; al colegio. Pero ¿quién hace la compra? Al final de la semana todo se acaba. No queda fruta los viernes. Los sábados dejaron la cuenta al cero. Y los domingos, aunque te dices que vivirás una feliz experiencia, la cosa no es tan sencilla: el niño sale con sus amigos. ¿Hay algún chico en la esquina? ¿Se cayó en el jardín? Desde luego, siempre gozan de perfecta salud y yo estoy aquí preocupándome por nada. Definitivamente, vivir no es tan sencillo ni al final del siglo veinte.

A.3.3.2 Frases

1) No queda fruta los viernes
2) ¿Ya se cambió de ropa?
3) ¿Hay algún chico en la esquina?
4) El final del siglo veinte.
5) ¿La puerta tiene cerrojo?

6) Tengo la llave en el bolsillo.
7) ¿Se cayó en el jardín?
8) ¿Rompió la yema del huevo?
9) Gozan de perfecta salud.
10) Vivirás una feliz experiencia.
11) Dejaron la deuda al cero
12) Le gusta mucho el gregoriano.
13) Yo llevo al niño en el coche.
14) Llegó la reina del puño cerrado.
15) Arrizabalaga dejará la reyerta.

A.3.3.3 Palabras

1) la yema
2) jardín
3) huevo
4) se cayó
5) la llave
6) el bolsillo
7) la puerta
8) cerrojo
9) veinte
10) el final
11) chico
12) esquina
13) ropa
14) se cambió
15) fruta
16) no queda
17) niño
18) coche
19) gregoriano
20) le gusta
21) la deuda
22) cero
23) experiencia

24) vivirás

25) gozan

26) reina

27) salud

28) llegó

29) cerrado

30) Arrizabalaga

31) reyerta

6.2.3.3 Relación entre las frases y las palabras de la base de datos

Número de frase	Frase portadora	Palabra o grupo aislados	Posición dentro de la frase
1	No queda fruta los viernes.	Fruta No queda	Media Inicial
2	¿Ya se cambió de ropa?	Se cambió Ropa	Media Final
3	¿Hay algún chico en la esquina?	Chico Esquina	Media Final
4	El final del siglo veinte.	El final veinte	Inicial Final
5	¿La puerta tiene cerrojo?	La puerta cerrojo	Inicial Final
6	Tengo la llave en el bolsillo.	Llave Bolsillo	Media Final
7	¿Se cayó en el jardín?	Se cayó jardín	Inicial Final
8	¿Rompió la yema del huevo?	La yema huevo	Media Final
9	Gozan de perfecta salud.	Gozan salud	Inicial Final
10	Vivirás una feliz experiencia.	Vivirás experiencia	Inicial Final
11	Dejaron la deuda al cero.	Deuda cero	Media Final
12	Le gusta mucho el gregoriano.	Le gusta gregoriano	Inicial Final
13	Yo llevo al niño en el coche.	Niño Coche	Media Final
14	Llegó la reina del puño cerrado.	Llegó Reina	Inicial Media
15	Arrizabalaga dejará la reyerta.	Arrizabalaga reyerta	Inicial final

A.3.4 Cuestionario de evaluación de voz emotiva en el proyecto VAESS

Identificación de la emoción en voz natural

Grabación	Neutra	Alegre	Triste	Enfadada	No Identificable	Otras
1						
2						
…						

19						
20						

Evaluación global

¿Cómo de natural le suena la voz con emociones?

__ La voz parece muy natural

__ La voz parece más bien natural

__ La voz parece más la de un robot que una voz natural

¿Cómo de inteligible es el habla?

__ buena __ aceptable __ pobre

¿Como juzga la calidad de la voz sintetizada?

__ buena __ aceptable __ pobre

Por favor comente libremente sus opiniones sobre la voz y las emociones

A.3.5 Cuestionario sobre la personalización de voz

Las frases empleadas han sido ...

Parámetros finales del proceso

¿De qué tipo de voz personalizada le gustaría disponer?

_____

¿Qué problemas encontró a lo largo del proceso?

_____

Por favor, comente con libertad su opinión acerca de la voz y su proceso de personalización

---

	Voz Personal				**Voz estándar**		
Frase	**Grado de similitud con lo deseado (1-7)**	**Naturalidad (1-7)**	**Inteligibilidad (1-7)**	**Aceptabilidad (1-7)**	**Naturalidad (1-7)**	**Inteligibilidad (1-7)**	**Aceptabilidad (1-7)**
1							
2							
3							
4							
5							

A.3.6 Definición de rasgos simples y complejos para la voz personalizada o con emociones

00.03 DEF <VOK>:=<+SEG,+FON,-DIG,+VOC,-CONS,-PUNKT>

```
00.04 DEF <VOKDEF>:=<+VOK,+VOICE,-NAS,+CONT,-SINTAX,-BLISS,-DIPH,-OBST,-HSTR,-
MSTR,-LSTR,-STRESS,-1STRESS,-TENSE,-FRIC,-DIPT>
00.05 DEF <KONS>:=<+SEG,+FON,-DIG,+CONS,-PUNKT>
00.06 DEF <KONSDEF>:=<+KONS,-VOC,-SINTAX,-BLISS,-DIPH,-HSTR,-MSTR,-LSTR,-STRESS,-
1STRESS,-TENSE,-DIPT>
  // clasificación de los fonemas:
  // - vocales: +SEG, +VOICE, +CONT, +VOC (FONEMAS, SONORAS, CONTINUAS, VOCÁLICAS)
  // - consonantes: +SEG, -VOC (FONEMAS, CONSONÁNTICOS)
  // * fricativas: +FRIC (USAN LA RAMA PARALELO)
  //      _ sonoras: +VOICE como /J/
  //      _ no sonoras: -VOICE
  //       - aspiradas: como 'S' 'F' 'Z'
  //       - no aspiradas: como 'X'
  //     * no fricativas: +FRIC (NO USAN LA RAMA PARALELO)
  //       _ continuas: +CONT
  //    _ no continuas: -CONT
00.07 DEF <SEMICON>:=<HSTR>
00.08 DEF <SEMIVOC>:=<MSTR>
00.09 DEF <SEMIKON>:=<+VOK,+SEMICON>
00.10 DEF <SEMIVOK>:=<+VOK,+SEMIVOC>
01.00 DEF <AUXILIA>:=<LSTR>
01.01 DEF <PALFUNC>:=<DIPH>
01.02 DEF <FUERTE>:=<-SEMIVOC,-SEMICON>
01.03 DEF <PUNTDEF>:=<WB,FB,-SEG,PUNKT,-DIG,+FON,STRESS,-BLISS,+SYNTAX>
01.04 DEF <SIGNDEF>:=<WB,FB,-SEG,PUNKT,-DIG,STRESS,-BLISS,-SYNTAX>
01.05 DEF <DIGDEF>:=<-WB,-FB,-SEG,-PUNKT,DIG,-FON,-BLISS,-SYNTAX>
01.06 DEF <HLB>:=<+HIGH,-LOW,+BACK>
01.07 DEF <OCLUDEF>:=<+KONSDEF,+OBST,-CONT,-NAS>
  desde el punto de vista de los formantes son más bien continuas
01.08 DEF <NASDEF>:=<+KONSDEF,-OBST,+CONT,+NAS,+VOICE,-FRIC>
01.09 DEF <AFRIDEF>:=<+KONSDEF,+OBST,-CONT,+FRIC,-NAS>
01.10 DEF <LATEDEF>:=<+KONSDEF,-OBST,+CONT,-NAS,+VOICE,-FRIC>
02.00 DEF <APRODEF>:=<+KONSDEF,+OBST,+CONT,-NAS,+VOICE,+FRIC>
02.01 DEF <FRICDEF>:=<+KONSDEF,+OBST,+CONT,-NAS,-VOICE,+FRIC>
02.02 DEF <VIBRDEF>:=<+KONSDEF,-OBST,-CONT,-NAS,+VOICE,-FRIC>
02.03 DEF <OCLUS>:=<+OCLUDEF>
//  vocales
02.04 DEF A:=<+VOKDEF,-HIGH,+LOW,+BACK,-ROUND>
02.05 DEF E:=<+VOKDEF,-HIGH,-LOW,-BACK,-ROUND>
02.06 DEF I:=<+VOKDEF,+HIGH,-LOW,-BACK,-ROUND>
02.07 DEF O:=<+VOKDEF,-HIGH,-LOW,+BACK,+ROUND>
02.08 DEF U:=<+VOKDEF,+HIGH,-LOW,+BACK,+ROUND>
//  OCLUSIVAS SONORAS
02.09 DEF B1:=<+OCLUDEF,+VOICE,-FRIC,+ANT,-COR>
02.10 DEF D1:=<+OCLUDEF,+VOICE,-FRIC,+ANT,+COR>
03.00 DEF G1:=<+OCLUDEF,+VOICE,-FRIC,-ANT,-COR,+HLB>
//  OCLUSIVAS SORDAS
03.01 DEF P:=<+OCLUDEF,-VOICE,+FRIC,+ANT,-COR>
03.02 DEF T:=<+OCLUDEF,-VOICE,+FRIC,+ANT,+COR>
03.03 DEF K:=<+OCLUDEF,-VOICE,+FRIC,-ANT,-COR,+HLB>
//  APROXIMANTES
03.04 DEF B:=<+APRODEF,+ANT,-COR>
03.05 DEF D:=<+APRODEF,+ANT,+COR>
03.06 DEF G:=<+APRODEF,-ANT,-COR,+HLB>
```

```
03.07 DEF J:=<+APRODEF,-ANT,-COR,+HLB>
//  FRICATIVAS SORDAS
03.08 DEF F:=<+FRICDEF,+ANT,-COR>
03.09 DEF S:=<+FRICDEF,+ANT,+COR>
03.10 DEF TH:=<+FRICDEF,+ANT,+COR>
04.00 DEF X:=<+FRICDEF,-ANT,-COR,+HLB>
  //  NASALES
04.01 DEF M:=<+NASDEF,+ANT,-COR>
04.02 DEF N:=<+NASDEF,+ANT,+COR>
04.03 DEF NY:=<+NASDEF,-ANT,+COR>
  //  VIBRANTES
04.04 DEF R:=<+VIBRDEF,-ANT,-COR,+HLB>
04.05 DEF R1:=<+VIBRDEF,-ANT,-COR,+HLB>
04.06 DEF RR:=<+VIBRDEF,-ANT,-COR,+HLB>
  //  LATERALES
04.07 DEF L:=<+LATEDEF,+ANT,+COR>
04.08 DEF L1:=<+LATEDEF,+ANT,+COR>
04.09 DEF LL:=<+LATEDEF,+ANT,+COR>
  //  AFRICADAS
  He puesto en FRIC + porque no tenía ningún signo; no se si es correcto, porque
es sólo a medias fric.
04.10 DEF CH:=<+AFRIDEF,+HIGH,-LOW,-BACK,-VOICE,-ANT,-COR>
  Definimos la 'J1' que se da en posición inicial absoluta o tras 'n' o 'l'.No
podemos emplear 'J1' al se la / un signo emplead en el lenguaje de las reglas.
05.00 DEF J1:=<+AFRIDEF,+HIGH,-LOW,-BACK,+VOICE,-ANT,-COR>
  //AUXILIAR
05.01 DEF Q:=<+KONSDEF,OBST,LOW,BACK,+CONT,-NAS,-VOICE,-FRIC>
05.02 DEF J2:=<+KONSDEF,OBST,LOW,BACK,+CONT,-NAS,-VOICE,-FRIC>
05.03 DEF C:=<+KONSDEF>
05.04 DEF Z:=<+KONSDEF>
05.05 DEF H:=<+KONSDEF>
05.06 DEF V:=<+KONSDEF>
05.07 DEF Y:=<+KONSDEF>
05.08 DEF x:=<+KONSDEF>
05.09 DEF y:=<+KONSDEF>
05.10 DEF z:=<+KONSDEF>
06.00 DEF W:=<+KONSDEF>
  // signos de puntuación...
06.01 DEF ~:=<-VOC,-CONS,+SEG,-DIG,-FON,-ACCENT,-BLISS,-SYNTAX>
06.02 DEF  :=<WB,FB,-SEG,-PUNKT,-DIG,+FON,STRESS,-BLISS,-SYNTAX>
06.03 DEF .:=<+PUNTDEF>
```

A.3.7 Reglas segmentales y de entonación para el castellano (para personalización y para emociones)

```
//  VOCALES
13.06 DEF A:=(TMIN 7,T 12,ST 0 48,CS 3 80,FS 3 80,A0 0 28,AH 0 0 0 0,AC 0 0,AN 3
0,FD -1 100,F1 -1 700,F2 -1 1300,F3 -1 2300,FN -4 250)
13.07 DEF E:=(TMIN 7,T 11,ST 0 48,CS 3 80,FS 3 80,A0 0 30,AH 0 0,AC 0 0,AN 3 0,FD
-1 100,F1 -1 450,F2 -1 1800,F3 -1 2500      ,FN -4 250)
   suavizamos el polo nasal
13.08 DEF I:=(TMIN 7,T 10,ST 0 48,CS 3 80,FS 3 80,A0 0 30,AH 0 0,AC 0 0,AN 3 0,FD
-1 100,F1 -1 290 0 290,F2 -1 2100,F3 -1 2500,FN -4 250)
13.09 DEF O:=(TMIN 7,T 11,ST 0 48,CS 3 80,FS 3 80,A0 0 30,AH 0 0,AC 0 0,AN 3 0,FD
-1 100,F1 -1 500,F2 -1 900,F3 -1 2500,FN -4 250)
```

13.10 DEF U:=(TMIN 7,T 12,ST 0 48,CS 3 80,FS 3 80,A0 0 30,AH 0 0,AC 0 0,AN 3 0,FD -1 100,F1 -1 290 0 290,F2 -1 700,F3 -1 2500,FN -4 250)

// **CONSONANTES // OCLUSIVAS SONORAS**

Subimos velocidad ST; subimos CS y FS; damos ganancia vocálica, A0. Quitamos AH; reducimos AC. Variamos los valores de F1,F2 y F3. Variamos de nuevo, subiendo F1 y bajando F2.Restituimos los valores originales. Bajamos F1 y F2. Restituimos momentáneamente 2 puntos en F2 y en las frecuencias del original. Ahora estaba en 800 en la trama -1

14.00 DEF B1:=(TMIN 5,T 6,ST 0 48,CS -1 200,FS 0 130,A0 0 20,AH 0 0 0 0,AC 0 0 1 0 5 0,AN 0 10 1 0,AK -3 255,FD -1 90,F1 -1 200,F2 -3 650,F3 -1 2200,K1 -3 850,K2 -3 1600,FN -1 200)

Quitamos AH ;le damos valores a A0; variamos los valores de fricación. Cambiamos los valores de F1,F2 y F3; eliminamos AN. Reducimos AC; bajamos a 8 la fricación. Cambiamos a valores m s bajos K1 y K2. Restituimos valores originales de ST, CS y FS(en 3 en vez de en -3. Hemos restituido el AN y hemos quitado la fricación. Damos sonoridad. Subimos f2.Aumentamos velocidades de cs y fs y retrasamos su punto de aplicación

14.01 DEF D1:=(TMIN 5,T 6,ST 0 48,CS -1 200,FS 1 200,A0 0 20,AH 0 0 0 0,AC 0 0 1 0 5 0,AN 0 30 1 0,AK -3 0,FD -1 90,F1 -1 264,F2 -1 1500,F3 -1 2594,K1 -3 1000,K2 -3 2000,FN -1 200)

Quitamos AH; damos valores a A0; bajamos AC; subimos F1. subimos F2. Damos sonoridad. aceleramos los formantes superiores

14.02 DEF G1:=(TMIN 5,T 6,ST 0 48,CS -4 130,FS -5 80,A0 0 20,AH 0 0 0 0,AC 0 0 1 0 5 0,AN 0 10 1 0,AK -3 255,FD -1 90,F1 -1 350,F2 -1 1600,F3 -1 1900,K1 -3 1650, K2 -3 2000,FN -1 300)

// **OCLUSIVAS SORDAS**

Quitamos la Aspiración; bajamos AC. Retrasamos el F2. Subimos el F2. Eliminamos el segundo F2. Restauramos el primitivo. Retrasamos los paralelo. bajamos ák' y adelantamos de nuevo los paralelos. Bajamos ác. Retrasamos CS

14.03 DEF P:=(TMIN 6,T 11,ST 0 48,CS 1 130,FS 0 130,A0 0 0,AH 0 0 1 0 2 0,AC 0 0 1 15 2 0,AN 0 0 1 0,AK -3 28,FD -1 110,F1 0 200,F2 0 800,F3 0 2300,K1 1 800,K2 1 1600)

Quitamos la aspiración .BAJAMOS K1 Y K2 Y AK LO MODIFICAMOS. ADELANTAMOS LA TRAMA DE aplicación DE K1 y K2. Retrasamos CS

14.04 DEF T:=(TMIN 6,T 11,ST 0 48,CS 1 130,FS 3 80,A0 0 0,AH 0 0 0 0,AC 0 0 1 16 2 0,AN 0 0 1 0,AK -3 28,FD -1 110,F1 0 200,F2 0 2000,F3 0 3000,K1 1 800,K2 1 1600)

Quitamos AH. devolvemos la velocidad original a ST. Subimos k2. Reducimos AC. Subimos 'F1' que ahora está en 200. Ahora AC' está así:0 0 1 8 2 0;cambiamos. Volvemos a lo de antes. Retrasamos CS y FS. Aceleramos FS. Restauro el valor de F1. Subimos F3. Retraso los formantes. Ponemos más fricación para la barra de explosión.. Había 8.Ponemos sólo 10 de fricación porque si no suena demasiado. Vamos a hacer una prueba poniendo fric. En dos momentos, tal como aparece la doble barra de 'k' en habla natural. No funciona con la doble barra. Suena a 'ch'. Cambiamos de momento K2

14.05 DEF K:=(TMIN 6,T 11,ST 0 48,CS 1 130,FS 1 130,A0 0 0,AH 0 0 0 0,AC 0 0 1 8 2 0,AN 0 0 1 0,AK -3 255,FD -1 90,F1 2 200,F2 2 1700,F3 2 2500,K1 3 1800,K2 3 4000)

// **FRICATIVAS SONORAS**

Quitamos AH y metemos AC. Dejamos sólo una trama para las frecuencias de F2.Restituimos las dos tramas para F2; quitamos de nuevo una. Bajamos el valor de F1.Restituimos su valor original a F1. Restauramos el doble formante F2 de las labiales. Eliminamos el segundo F2. Incorporamos los formantes paralelo. Corregimos colocación de la fricación y aumentamos de la 'supuesta' 5 a 10.

14.06 DEF B:=(TMIN 5,T 6,ST 0 48,CS 3 80,FS 3 60,A0 0 14,AH 0 0,AC 0 0,AN 4 0,FD -1 95,F1 0 250,F2 -1 800,F3 -1 2150,K1 -3 1000,K2 -3 2500)

Quitamos AH y metemos AC; quitamos K1 y K2;cambiamos AK; quitamos AK . Subimos la frecuencia de F1; ponemos F1 y F2 en la trama 0. Restituimos su valor original a F1. Restituimos la serie paralelo. Corregimos error de posición de la fricación que estaba mal situada, por lo cual no había tal fricación.

14.07 DEF D:=(TMIN 5,T 6,ST 0 48,CS 3 80,FS 3 60,A0 0 14,AH 0 0,AC 0 0,AN 0 0,FD -1 95,F1 0 200,F2 0 1900,F3 -1 2900,K1 -3 1000,K2 -3 3000)

Quitamos AH; ponemos AC: subimos A0; subimos F1; Bajamos el valor de F1 que lo pusimos demasiado alto. Subimos los F2 y F3. Cambiamos el comienzo y velocidad de FS. Eliminamos AF, la barra de fricación. En las tres aproximantes. Reducimos TMIN. Corregimos el mismo error de las anteriores.

14.08 DEF G:=(TMIN 5,T 6,ST 0 48,CS 3 80,FS 3 60,A0 0 14,AH 0 0,AC 0 0,AN 0 0,FD -1 110,F1 -1 400,F2 -1 1500,F3 -1 2000,K1 -3 1000,K2 -3 2000,B1 0 40 1 0)

Metemos fricación. Hemos variado, subiendo, las velocidades de las transiciones de todos los formantes y en CS y FS las hemos adelantado . Hemos rebajado la ganancia vocálica. Subimos AC. Hemos bajado el valor de la fricación y hemos bajado la frecuencia de F1. bajamos K1 y K2; bajamos más la fricación. suavizamos el polo nasal. Hemos restituido k1 y k2 a sus valores originales. ponemos 'fs' en -1 y bajamos la velocidad. Bajamos sonoridad. BAJAMOS fricación, PONIENDO sólo UN PUNTO DE aplicación. bajamos 'k1' y 'k2'. Quitamos el 'fn', que está en 250 y el án que está en 4 0. Bajamos F2 de 2000 a 1800 para probar y variamos K1 de 1000 a 2000 y K2 de 2000 a 6000. Quitamos la mitad de fricación. Quitamos la fricación.

14.09 DEF J:=(TMIN 5,T 11,ST 0 28,CS -1 60,FS -1 60,A0 0 18,AH 0 0 ,AC 0 0,AK -3 112,FD -1 95,F1 -1 250,F2 -1 1800,F3 -1 3200,K1 -3 2000,K2 -3 6000)

// FRICATIVAS SORDAS

Eliminamos los formantes serie. Eliminamos aspiración y aumentamos fricación. Hemos modificado los formantes paralelos. ADELANTAMOS EL PUNTO DE aplicación DE LA transición A LA 'S'. Eliminamos f1 f2. Restituimos f1 f2. Aumentamos ac. Modificamos k1 k2. Eliminamos aspirac y f1 f2 f3 f4. Reducimos ac. Retocamos los formantes paralelos. ELIMINAMOS LA aspiración. AUMENTAMOS LA fricación. Eliminamos f1, f2, adelantamos ac. Restauramos f1 f2, aumentamos ac. bajamos K1. Estrechamos anchos de banda paralelo. ensanchamos un poco más. Restituimos los valores originales de K1, K2.bajar K2 (=5800) .Con esto mejora y no hay ploc. Eliminar AH (=0)Hecho. Me gusta como queda; creo que se puede quedar AC como está. Subir AC (=20).Probamos de todas formas a poner m s AC. No nos gusta, la dejamos como antes. ensanchar C1 C2 (C1 -3 60,C2 -3 60,) Hecho. Eliminamos C2. Subimos C1. Eliminar B1, B2 y la regla asociada en Fon2.sp. bajamos K1 y K2. Subimos de nuevo K1 y K2. Jugamos con AK, subiéndolo. El resultado es que anula K1. Subimos m s AK. El resultado acústico es indiferenciable. Restituimos AK. AK=128. adelantamos AC. Adelantamos AC pero por regla. Fijamos C2. Subimos a 15 AC. Hemos bajado AK y hemos variado C1 y C2.Otra vez el AK a su sitio!.. Subimos K2 y ponemos C1 y C2 en 200. Bajamos un poquito á0'.BAJAMOS DE 12 LA FRICACIÓN PARA VER EL EFECTO.

14.10 DEF S:=(TMIN 7,T 13,ST 0 48,CS 3 80,FS 3 80,A0 0 0,AH 0 0,AC 0 8,AN 0 0,AK -3 32,FD -1 105,F1 -1 200,F2 -1 1900,F3 -1 2900,K1 -3 4000,K2 -3 6000,C1 -3 40,C2 -3 60)

Cambiamos AH y AC .Modificamos K2,bajamos FS, HEMOS RETRASADO LA TRAMA DE aplicación DE LOS FORMANTES SERIE. Restituimos en F2 la trama de aplicación del formante. Ponemos K1 al máximo y quitamos K2 y metemos el máximo de fricación. Reducimos un poco la fricación, pero parece que, efectivamente, el 'ploc' desaparece al quitar K2.Hemos bajado un poco la fricación. Probamos a

quitar toda la aspiración. Si no hay asp no hay f1..f4. Ponemos K2 Restituimos los formantes serie. más anchos los formantes paralelo (C1 -3 200, C2 -3 200). Eliminar B1 y la regla asociada en Fon2.sp. Bajamos de 20 a 15 AC.

15.00 DEF F:=(TMIN 7,T 12,ST 0 48,CS 0 130,FS 0 80,A0 0 0,AH 0 0 ,AC 0 15,AN 0 0, AK -3 160,FD -1 105,F1 -1 250,F2 -3 950 0 950,F3 -1 1900,K1 -3 850,K2 -3 6000,C1 0 200,C2 0 200)

Bajamos AH y subimos AC y adelantamos a 0 la AC: Ponemos 2 puntos de aplicación de AH. Reducimos la aspiración. Eliminamos los formantes serie, que son copiados de la vocal siguiente o de la anterior. Modificamos los formantes paralelo. Incrementamos la velocidad de transición de formantes (FS, CS). Modificamos la trayectoria de AH y hemos aumentado su velocidad de transición ST. Probamos con AH a 0 y subimos AC. Subimos un poco m s AC y ponemos un poco de AH. Bajamos AH. Subimos AC. Cambiamos el valor de AK. Subimos ST. C1 era ancho. Ak no debería tender a cero. LA transición DE AC DEBE SER más SUAVE. PONEMOS ac IGUAL EN LOS DOS PUNTOS Y SUBIMOS AH. Abrimos c2. Subimos al máximo la fricación. bajamos y 1 punto. Ák' QUE TIENE AHORA 28 EN -3 LO PONEMOS A 0.

15.01 DEF X:=(TMIN 7,T 14,ST 0 48,CS 3 80,FS -5 50,A0 0 0,AH 0 20 ,AC 1 10,AN 0 0, AK -3 0,FD -1 95,F1 -1 500,F2 -1 1300 ,F3 -1 2400,F4 -1 3200,K1 -3 1000,K2 -3 2000,B1 0 40 1 0)

Se eliminan los formantes serie . Se elimina aspiración y aumentamos la fricación. Reducimos la fricación y restauramos los formantes serie. Eliminamos los c1 c2. Cambiamos AK. Restituimos anchos de banda paralelos y quitamos B1. Hemos rebajado los anchos de banda paralelo. KA lo hemos puesto a 128 y hemos bajado K1.

15.02 DEF TH:=(TMIN 6,T 12,ST 0 64,CS 0 80,FS 0 80,A0 0 0,AH 0 0,AC -1 14,AN 0 0, AK -3 128,FD -1 105,F1 -1 200,F2 -1 1400,F3 -1 2600,K1 -3 1800,C1 -3 40,K2 -3 4900,C2 -3 60)

*// **NASALES***

Bajamos la ganancia vocálica . Subimos FS; cambiamos el valor de F1 y F2, F3, F4. Retrasamos F2. Reducimos la nasalidad. Retrasamos el momento de aplicación de AN. Retrasamos también. El momento de aplicación de B1. Devolvemos la velocidad inicial de 'CS'. Rebajamos y retrasamos nasalidad y bajamos ganancia vocálica.

15.03 DEF M:=(TMIN 5,T 9,ST 0 48,CS 0 130,FS 0 130,A0 0 15,AH 0 0,AC 0 0,AN -2 18, FD -1 95,F1 -1 275,F2 -1 900 0 900,F3 -1 2100,B1 -3 40 5 0,B2 -2 48 1 0)

bajamos la ganancia vocálica. Retrasamos el momento de aplicación de AN y B1. Retrasamos una trama la nasalidad y bajamos ésta y la ganancia vocálica (ahora -3 20, 18)

15.04 DEF N:=(TMIN 5,T 13,ST 0 48,CS 3 80,FS 3 80,A0 0 15,AH 0 0,AC 0 0,AN -2 18, FD -1 95,F1 -1 300,F2 -1 1600,F3 -1 2600,B1 -3 40 5 0,B2 -2 48 1 0)

Bajamos A0. Cambiamos los valores de los formantes serie. Subimos la velocidad de las transiciones de todos los formantes. Hemos subido F1. Adelantamos AN y B1. Bajamos FS y CS. Hemos puesto los F serie con los valores del tonto. Bajamos A0.Retrasamos y bajamos nasalidad.

15.05 DEF NY:=(TMIN 5,T 13,ST 0 48,CS 3 80,FS 3 80,A0 0 15,AH 0 0,AC 0 0,AN -2 18, FD -1 105,F1 -1 400,F2 -1 1900,F3 -1 2600,B1 -3 40 5 0,B2 -2 48 1 0)

*// **LATERALES***

Cambiamos frecuencias de Formantes 1,2 y 3 (según Quilis). Bajamos ganancia vocálica; bajamos velocidad de FS; subimos FS. Subimos CS. Hemos disminuido la duración de 'L' porque nos parecía muy larga. Subimos A0. Restauramos los formantes. Suavizamos el polo nasal

15.06 DEF L:=(TMIN 5,T 10,ST 0 48,CS -4 180,FS 3 100,A0 0 25,AH 0 0,AC 0 0,AN 4 0, FD -1 95,F1 -1 275,F2 -1 1300,F3 -1 3000,FN -1 250)

Definimos un tipo diferente de 'l' cuando va precedida de 'k' o de 'g'. Aumentamos A0 y los valores de los formantes serie a los valores de la 'l' inicialmente definida por los suecos. Hemos disminuido la duración. Bajamos f1 y suavizamos el polo nasal. b jamos A0 de 25 a 18.

15.07 DEF L1:=(TMIN 5,T 12,ST 0 48,CS 0 180,FS 0 130,A0 0 18,AH 0 0,AC 0 0,AN 4 0, FD -1 95,F1 -1 300,F2 -1 1300,F3 -1 2564 ,FN -1 250)

Este fonema es de nueva creación, ya que los suecos lo tenían en el mismo paquete que la fricativa 'J'. Retocamos un poco los valores de formantes y las velocidades de transición y aumentamos su duración intrínseca. Suavizamos el polo nasal

15.08 DEF LL:=(TMIN 5,T 12,ST 0 60,CS -4 140,FS 3 100,A0 0 25,AH 0 0,AC 0 0,AN 4 0,FD -1 95,F1 -1 250,F2 -1 2085,F3 -1 2766 ,FN -1 250)

// VIBRANTES

Cambiamos la duración de 'r' según los valores de Quilis que son 20 mms.

15.10 DEF R:=(TMIN 5,T 20,ST 0 48,CS 3 80,AK -3 0,FS 3 80,A0 0 30 1 0 2 30,AH 0 0, AC 0 0,AN 0 0,FD -1 95,F1 -1 500,F2 -1 1400,F3 -1 2300,K1 -3 800,K2 -3 1600)

Cambiamos la ganancia vocálica, A0. Cambiamos, subiéndola la velocidad del formante rápido, CS. Cambiamos ST, fuente del tiempo, Bajándolo, Cambiamos, subiéndolo, FS, formante rápido para F2. Eliminamos los formantes serie porque dependen del contexto. Hemos cambiado los valores de la ganancia vocálica. Devolvemos su valor inicial a 'st'y a 'cs'y a 'fs'. Subimos 'ST'. redistribuimos á0'. Aumentamos CS y FS

16.00 DEF RR:=(TMIN 5,T 25,ST 0 28,CS 0 250,AK -3 0,FS 0 160,A0 0 0 1 0 3 20 4 20 5 0 6 0 7 20 8 20 9 0,F1 0 200 3 500,AH 0 0,AC 0 0,AN 0 0,FD -1 95)

// AFRICADAS

Dejamos AH A0 y ponemos valores más altos en AC. Hemos cambiado los valores de K1 y K2 y sólo jugamos con la fricación. Vamos a meter un poco de AH. Quitamos la aspiración porque suena peor. BAJAMOS áC'. Eliminamos áC' inicial. Retrasamos los formantes. Ahora tenemos 2 momentos de AC 0 0 5 18. Ponemos 2 momentos con AC, a ver qué pasa. Restituimos lo anterior con un poquito menos de fricación.

16.01 DEF CH:=(TMIN 7,T 13,ST 0 48,CS 3 80,FS 3 80,A0 0 0,AH 0 0 0 0,AC 0 0 5 16,AN 0 0,AK -3 0,FD -1 105,F1 0 250,F2 0 2479,F3 0 3100,K1 -3 2300,K2 -3 3000, B1 0 40 1 0)

Hemos definido una 'J1' africada. Suavizamos el polo nasal. BAJAMOS fricación A UN PUNTO. QUITAMOS án' Y 'FN'. Cambiamos 'k1' y 'k2',que estaban en 800 y 2000.Bajamos ganancia vocálica y fricación y subimos 'K2'de 3000 a 6000, como en 'J'. Bajamos también F2, de 2150. Aumentamos la duración de 10 a 16 que es valor de Quilis. Subimos 'CS' de 40.

16.02 DEF J1:=(TMIN 5,T 16,ST 0 48,CS 3 80,FS 3 40,A0 0 15,AH 0 0,AC 0 5,AK -3 0,FD -1 95,F1 -1 250,F2 -1 1800,F3 -1 3200,K1 -3 2000,K2 -3 6000)

Cambio los formantes paralelo de 2500 a 3900 y de 4000 a 4900. AÑADO ANCHOS DE BANDA paralelo: +200. Elimino anchos de banda paralelo: +200 para los signos de puntuación y pausas. Modificamos la duración intrínseca T. ACELERAMOS Y RETRASAMOS LA transición DE LAS GANANCIAS

16.03 DEF :=(F0SLOPE 25,A 0,B 0,C 0,D 0,E 0,RAS 0,T 10,ST -1 40,CS 3 80,FS 6 80,A0 0 0,AH 0 0,AC 0 0,AN 0 0,FD -1 100,F1 0 500,F2 0 1500,F3 0 2500,F4 0 3600,FH 0 4000,K1 0 4000,K2 0 4800,AK 0 128,F0 0 110 1 100,FN 0 206)

Añadimos un FN ficticio, eliminamos DR, modificamos la duración intrínseca T

16.04 DEF .:=(A 0,B 0,T 10,ST -8 40,CS 3 80,FS -1 80,A0 0 0,AH 0 0,AC 0 0,AN 0 0,FD -1 90,F1 0 500,F2 0 1500,F3 0 2500,F4 0 3600,K1 0 4000,K2 0 4800,AK 0 128, F0 5 95,FN 0 206)

```
16.05 DEF ,:=(A 0,B 0,T 10,ST -8 40,CS 3 80,FS -1 80,A0 0 0,AH 0 0,AC 0 0,AN 0
0,FD -1 90,F1 0 500,F2 0 1500,F3 0 2500,F4 0 3600,K1 0 4000,K2 0 4800,AK 0 128,
F0 5 120,FN 0 206)
      Defino : y ; con los mismos valores que ,
16.06 DEF ::=(A 0,B 0,T 10,ST -8 40,CS 3 80,FS -1 80,A0 0 0,AH 0 0,AC 0 0,AN 0
0,FD -1 90,F1 0 500,F2 0 1500,F3 0 2500,F4 0 3600,K1 0 4000,K2 0 4800,AK 0 128,
F0 5 120,FN 0 206)
16.07 DEF ;:=(A 0,B 0,T 10,ST -8 40,CS 3 80,FS -1 80,A0 0 0,AH 0 0,AC 0 0,AN 0
0,FD -1 90,F1 0 500,F2 0 1500,F3 0 2500,F4 0 3600,K1 0 4000,K2 0 4800,AK 0 128,
F0 5 120,FN 0 206)
      // SIGNOS DE PUNTUACIÓN
16.08 DEF ?:=(A 0,B 0,T 10,ST -8 40,CS 3 80,FS -1 80,A0 0 0,AH 0 0,AC 0 0,AN 0
0,FD -1 90,F1 0 500,F2 0 1500,F3 0 2500,F4 0 3600,K1 0 4000,K2 0 4800,AK 0 128,
F0 0 130 6 110,FN 0 206)
16.09 DEF !:=(A 0,B 0,T 10,ST -8 40,CS 3 80,FS -1 80,A0 0 0,AH 0 0,AC 0 0,AN 0
0,FD -1 90,F1 0 500,F2 0 1500,F3 0 2500,F4 0 3600,K1 0 4000,K2 0 4800,AK 0 128,
F0 5 120 6 110,FN 0 206)
16.10 DEF +:=(DR 0 50,A0 0 0 48 0,AH 0 0,AC 0 0,AN 0 0)
//REGLAS CONTEXTUALES
10.01:   ^ / &   ;$
      borra los '@' que insertó "DIG.SP"
10.02: @ ^
10.03:   ^ <-STRESS> / <> &
10.04 LCRULE   ^ / <-SEG> &
10.05 LCRULE   ^ / <> & <-SEG>se borra la coma tas punto
10.06: , ^ / & ,(,) .
      las palabras función se acentúan en final de frase o delante de pausa
10.07: + ^ / & <-SEG,PUNKT,STRESS>
10.08: ' ^ / & <SEG>(,) +
      Una vez borradas las comillas, debemos corregir la conversión grafema-
fonema (debido a la existencia de palabras función nos vemos obligados a
repetir la conversión contemplada en el fichero GRAF.SP) 'U' sin acentuar es
la semiconsonante 'W' delante de vocal
10.09: <U,-STRESS> ^ <+SEMICON> / & <VOK>
      'I' sin acentuar es la semiconsonante delante de vocal
10.10: <I,-STRESS> ^ <+SEMICON> / & <VOK>
      'U' sin acentuar es semivocal tras vocal
11.00: <U,-STRESS> ^ <+SEMIVOC> / <VOK> &
      'I' sin acentuar es semivocal tras vocal
11.01: <I,-STRESS> ^ <+SEMIVOC> / <VOK> &
11.02: <VOK,-STRESS> ^ <STRESS,1STRESS> / ' &
11.03: <VOK,-STRESS> ^ <STRESS> / ' &
11.04: ' ^
11.05: <KONS,VOICE,-CONT,-NAS> ^ <AN=0> / & <KONS,-VOICE>
      se borran los espacios tras palabras función
11.06:   ^ / + &
se borran los marcadores de palabras función '+' y se convierten en el rasgo
PALFUNC
11.07: <> ^ <-PALFUNC>
11.08: + ^ <+PALFUNC>
11.09: + ^   ;
      se borran los '#'
11.10: # ^
      'RR' es 'R' antes de no-fonema-no-signo-de-puntuación
```

12.00: RR ^ R **/ & <-SEG,-PUNKT>**
'RR' es 'R' en final de palabra
12.01: RR ^ R **/ & <-SEG,+PUNKT>**
corrección de la no-oclusividad de 'B' 'D' y 'G' en comienzo de palabra B significa, temporalmente, <-SEG> y comienzo absoluto de oración o tras pausa o nasal
12.02: <SEG> ^ <B:=0>
espacio inicial
12.03: <-SEG> ^ <B:=1>
separación entre palabras
12.04: <-SEG,-PUNKT> ^ <B:=0> **/ <> &**
'B' y 'G' son oclusivas tras nasal
12.05: <-SEG,-PUNKT> ^ <B:=1> **/ <+NASDEF> & <OCLUDEF>**
12.06: B1 ^ B **/ <-SEG,B=0> &**
12.07: G1 ^ G **/ <-SEG,B=0> &**
'B' es oclusiva tras 'L'
12.08: <-SEG,-PUNKT> ^ <B:=1> **/ L & D1**
'J' y 'D' no son oclusivas en comienzo de palabra
12.09 J1 ^ J **/ <-SEG,B=0> &**
12.10: D1 ^ D **/ <-SEG,B=0> &**
Calculamos la posición en la palabra, empleando el rasgo DIPT. Para los diptongos y triptongos, la primera vocal es <-DIPT>; y las demás son <+DIPT el número de sílabas por defecto es 0
13.00: <> ^ <B:=0>
13.01: <-SEG> ^ <-DIPT>
la primera vocal pertenece a la primera sílaba
13.02: <+VOK,-DIPT> ^ <B:=X+1> / <-SEG,X :=B> <+KONS>(,) &
una sílaba <-DIPT> no está en la mima sílaba que la anterior <-CONS><-DIPT>
13.03: <+VOK,-DIPT> ^ <B:=X+1> **/ <+VOK,-DIPT,X:=B> <+VOK,+DIPT>(,)** *<+KONS>(,) &*
una vocal <DIPT> está en la misma sílaba que la vocal previa
13.04: <+VOK,+DIPT> ^ <B:=X> **/ <X:=B,+VOK,-DIPT> &**
13.05: <+VOK,+DIPT> ^ <B:=X> **/ <X:=B,+VOK,+DIPT> &**
asignación temporal
13.06: <+KONS> ^ <B:=0>
una sílaba pertenece a la misma sílaba que la siguiente vocal
13.07: <+KONS> ^ <B:=X> **/ & <+VOK,X:=B>**
un grupo consonántico pertenece a la misma sílaba que la siguiente vocal
13.08: P ^ <B:=X> **/ & R1 <+VOK,X:=B>**
13.09: P ^ <B:=X> **/ & L1 <+VOK,X:=B>**
13.10: K ^ <B:=X> **/ & R1 <+VOK,X:=B>**
14.00: K ^ <B:=X> **/ & L1 <+VOK,X:=B>**
14.01: G ^ <B:=X> **/ & R1 <+VOK,X:=B>**
14.02: G ^ <B:=X> **/ & L1 <+VOK,X:=B>**
14.03: G1 ^ <B:=X> **/ & R1 <+VOK,X:=B>**
14.04: G1 ^ <B:=X> **/ & L1 <+VOK,X:=B>**
14.05: F ^ <B:=X> **/ & R1 <+VOK,X:=B>**
14.06: F ^ <B:=X> **/ & L1 <+VOK,X:=B>**
14.07: D ^ <B:=X> **/ & R1 <+VOK,X:=B>**
14.08: T ^ <B:=X> **/ & R1 <+VOK,X:=B>**
14.09: B ^ <B:=X> **/ & R1 <+VOK,X:=B>**
14.10: B1 ^ <B:=X> **/ & R1 <+VOK,X:=B>**

15.00: D1 ^ <B:=X> **/ & R1 <+VOK,X:=B>**
15.01: B ^ <B:=X> **/ & L1 <+VOK,X:=B>**
15.02: B1 ^ <B:=X> **/ & L1 <+VOK,X:=B>**
15.03: D1 ^ <B:=X> **/ & L1 <+VOK,X:=B>**
una consonante que no forma parte de un grupo consonántico o que está ligada a la siguiente consonante no está en la misma sílaba que la siguiente vocal, sino en la misma que la vocal previa
15.04: <+KONS,B=0> ^ <B:=X-1> **/ & <+KONS>(,) <VOK,X:=B>**
15.05: <+KONS> ^ <B:=X> **/ <X:=B> & <+KONS>(,) <-SEG>**
// el número de sílabas por defecto es 0
15.06: <> ^ <A:=0>
en los espacios en blanco copiamos el número de sílaba desde el último fonema
15.07: <-SEG> ^ <A:=Y> **/ <+SEG,Y:=B> &**
every phoneme copies the syllable number from the end of the word
15.08: <+SEG> ^ <A:=X> **/ & <+SEG>(,) <-SEG,X:=A>**
non-segments belong to no word
15.09: <-SEG> ^ <A:=0>
REGLAS DE DURACIONES omitidas por ser en general anteriores a esta Tesis
se eliminan todos lo espacios en blancos salvo el primero
34.05 LCRULE ^ **/ <-SEG> <+SEG>(,) &**
// ENTONACIÓN
se borran las variables temporales
34.06: <> ^ <A:=0>
34.07: <> ^ <B:=0>
34.08: <> ^ <C:=0>
34.09: <> ^ <D:=0>
34.10: <> ^ <E:=0>
35.00: <> ^ <RAS:=0>
35.01: <> ^ <-AUXILIA>
tiempo acumulado en t
35.02: <> ^ <T=0>
35.03: <> ^ <T=Z+Y> **/ <Z:=T,Y:=DR> &**
se asignan 110 Hz a las vocales acentuadas
35.04: <VOK,STRESS> ^ <F0:=110,TF0=0>
tras pausa se adelanta el pico de F0
35.05: <VOK,STRESS> ^ <TF0=-8> **/ <-SEG> <SEG,-STRESS>(,) &**
35.06: <STRESS,-CONS> ^ <LF0=X,A=Y+W-T> **/ & <-STRESS>(,) <W:=TF0,X:=F0,Y:=T>**
si la distancia es mayor que 125 insertamos nuevos puntos intermedios para superar la limitación de no poder interpolar más de 128 puntos
35.07: <X:=F0,A>125> ^ <A=125>
35.08: <X:=F0> ^ <LTF0=A>
*35.09: <VOK,STRESS> ^ <F0:=94+42*X2/100,TF0=DR+4>*
adelantamos el punto de asignación f0 tras pausa
35.10: <VOK,STRESS> ^ <TF0=DR-6> **/ & <SEG,-STRESS>(,) <-SEG>**
36.00: <STRESS> ^ <T=T.3>* **/ & <-STRESS>(,) ,**
36.01: <STRESS> ^ <T=T.6>* **/ , <-STRESS>(,) &**
añadimos la declinación
36.02: <> ^ <F0SLOPE:=25+X4-100>
// we lift contour 10 Hz and tilt it so the end will come
// F0SLOPE Hz lower
*36.03: <X:=F0> ^ <LF0= LF0 - F0SLOPE * (T+LTF0) / (Z+10), F0=X-F0SLOPE * (T+TF0) / (Z+10)>*
AUXILIA significa ahora "primera vocal acentuada"

```
 36.04: <> ^ <-AUXILIA>
36.05: <+VOK,+STRESS> ^ <+AUXILIA>
36.06 LCRULE <+VOK,+STRESS> ^ <-AUXILIA> / <+AUXILIA> <-AUXILIA>(,) &
detectamos si el primer punto de f0 está más allá de la trama 127
36.07: <+VOK,-STRESS> ^ <F0:=48+42*X2/100> / <-SEG> <+KONS>(,) & <-STRESS>(,) <
+STRESS,+AUXILIA,T>127>
      //  AJUSTES FONÉTICOS CONTEXTUALES
      borramos las variables auxiliares
36.08: <> ^ <A:=0>
36.09: <> ^ <B:=0>
36.10: <> ^ <C:=0>
37.00: <> ^ <RAS:=0>
37.01: <> ^ <-AUXILIA>
      Posición de la explosión y de la sonoridad 'CH' se compone de una pausas
y fricación  retasamos la fricación
37.02: CH ^ <TLAH=3*DR/4,TLAC=3*DR/4>
37.03: J1 ^ <TLAH=3*DR/4,TLAC=3*DR/4>
      no hay barra de sonoridad tras sorda o pausa: 'B1', 'D1', 'G1'
37.04: <VOICE,-CONT,-NAS,KONS> ^ <AN=0> / <-VOICE> &
      //  NASALIZACION
      las nasales no nasalizan el fonema sordo previo
37.05: <KONS,NAS> ^ <TAN=0> / <-VOICE> &
37.06: <X:=B1> ^ <LTB1=DR+LTB1-2>
37.07: <KONS,NAS> ^ <TLB1=X+DR-2> / & <X:=DR,VOICE,SEG>
37.08: <KONS,NAS> ^ <LTB2=DR-1>
      sonora después de nasal: AN vale 0 al principio de la sonora
37.09: <SEG,VOICE> ^ <TAN=0> / <KONS,NAS> &
      sonora después de nasal: AN vale 0 al final de la nasal
37.10: <KONS,NAS> ^ <TAN=DR> / <SEG,VOICE> &
      insertamos menos nasalización en el comienzo de la siguiente vocal con 2
tramas de solapamiento
38.00: <SEG,VOICE> ^ <AN:=15,1TAN=-2> / <KONS,NAS> &
38.01: <KONS,NAS> ^ <AN:=15,1TAN=-2> / <SEG,VOICE> &
      no son formantes, sino índices a tablas de formantes. limitamos los
valores máximos de F2 y F3
38.02: <F2>220> ^ <F2=220,F3=F2-78>
38.03: <F3>160> ^ <F3=160>
       'X' adopta los formantes de su contexto
38.04: X ^ <F1:=W,F2:=X,F3:=Y> / & <W:=F1,X:=F2,Y:=F3>
tras un no-segmento, copiamos los formantes de siguiente fonema
38.05: X ^ <F1:=W,F2:=X,F3:=Y> / <-SEG> & <W:=F1,X:=F2,Y:=F3>
      si va seguido por vocal y no va precedido por pausa,
copiamos los formantes de la vocal siguiente
38.06: X ^ <F1:=W,F2:=X,F3:=Y> / <+SEG> & <+VOK,W:=F1,X:=F2,Y:=F3>
      si va precedido por segmento y seguido por no pausa. Copiamos el
formante del fonema previo
38.07: X ^ <F1:=W,F2:=X,F3:=Y> / <+SEG,W:=F1,X:=F2,Y:=F3> & <-SEG>
      1st FORMANTE PARAL is not as narrow in 'XO' or 'XU'
38.08: X ^ <K1=0,C1=200,C2=0,TK2=-5,TC2=-5> / & <+ROUND>
38.09: <-FRIC> ^ <K1:=X, C1:=Y, K2:=W, C2:=Z, TK1=DR/3, TC1=DR/3, TK2=DR/3,
TC2=DR/3> / & <X, X:=K1, Y:=C1, W:=K2,Z:=C2> <+ROUND>
38.10: X ^ <K1=0,C1=200,C2=0,TK2=-5,TC2=-5> / & I
39.00: <-FRIC> ^ <K1:=X, C1:=Y, K2:=W, C2:=Z, TK1=DR/3, TC1=DR/3, TK2=DR/3,
TC2=DR/3> / & <X,X:=K1,Y:=C1,W:=K2,Z:=C2> I
```

si va seguido por pausa pero precedido por no pausa, copiamos los formantes del fonema anterior

39.01: RR ^ <LF1=X,F2:=Y,F3:=W> / **<+SEG,X:=F1,Y:=F2,W:=F3> & <-SEG>**

39.02: RR ^ <LF1=X,F2:=Y,F3:=W> / **& <+SEG,X:=F1,Y:=F2,W:=F3>**

modificamos contextualmente los formantes F1=368,F2=1234

39.03: RR ^ <LF1=84,F2=124> / **I &**

F1=462,F2=1165

39.04: RR ^ <LF1=116,F2=116> / **E &**

F1=550,F2=1199

39.05: RR ^ <LF1=140,F2=120> / **A &**

F1=412,F2=1068

39.06: RR ^ <LF1=100,F2=104> / **O &**

F1=336,F2=951

39.07: RR ^ <LF1=72,F2=88> / **U &**

F1=368,F2=1234

39.08: RR ^ <LF1=84,F2=124> / **& I**

F1=462,F2=1165

39.09: RR ^ <LF1=116,F2=116> / **& E**

F1=550,F2=1199

39.10: RR ^ <LF1=140,F2=120> / **& A**

F1=412,F2=1068

: RR ^ <LF1=100,F2=104> / **& O**

F1=336,F2=951

40.00: RR ^ <LF1=72,F2=88> / **& U**

F1=368,F2=1234

40.01: RR ^ <LF1=84,F2=124> / **& <-SEG,-PUNKT> I**

F1=462,F2=1165

40.02: RR ^ <LF1=116,F2=116> / **& <-SEG,-PUNKT> E**

F1=550,F2=1199

40.03: RR ^ <LF1=140,F2=120> / **& <-SEG,-PUNKT> A**

F1=412,F2=1068

40.04: RR ^ <LF1=100,F2=104> / **& <-SEG,-PUNKT> O**

F1=336,F2=951

40.05: RR ^ <LF1=72,F2=88> / **& <-SEG,-PUNKT> U**

asignamos los formantes de la 'R' final

40.06: RR ^ <LF1=#> / **<+VOK> & <+PUNKT>**

40.07: RR ^ <F1=84,F2=124> / **I & <+PUNKT>**

40.08: RR ^ <F1=116,F2=116> / **E & <+PUNKT>**

40.09: RR ^ <F1=140,F2=120> / **A & <+PUNKT>**

40.10: RR ^ <F1=100,F2=104> / **O & <+PUNKT>**

41.00: RR ^ <F1=72,F2=88> / **U & <+PUNKT>**

las transiciones entre continuas sonoras comienzan en la primera y son más bien lentas

X/**4,CS=130,TFS=-X/4,FS=80> / <SEG,VOICE,CONT,X:=DR> &**

diferenciamos entre continuas sordas y sonoras

41.01: <SEG,+VOICE,CONT> ^ <TF1=-X/4,TF2=-X/4,TCS=-X/4,CS=130,TFS=-X/4, FS=130> / <SEG,VOICE,CONT,X:=DR> &

41.02: <SEG,+VOICE,CONT> ^ <TF1=-X/4,TF2=-X/4,TCS=-X/4,CS=130,TFS=-X/4, FS=130> **/ <SEG,-VOICE,CONT,X:=DR> &**

41.03: <SEG,-VOICE,CONT> ^ <TF1=-X/4,TF2=-X/4,TCS=-X/4,CS=60,TFS=-X/4,FS=80> / **<SEG,CONT,X:=DR> &**

AUXILIA equivale a ser una 'X'

41.04: <> ^ <-AUXILIA>

```
41.05: X ^ <+AUXILIA>
      reglas contextuales para las sordas continuas que no sean 'X'
41.06: <SEG,-VOICE,CONT,-AUXILIA> ^ <F1=56> / & <+VOK,-HIGH>
      mejoramos la definición de las vocales delante de 'L' o 'LL'
41.07: <VOK> ^ <CS=130,FS=130> / <APRODEF> &
41.08: <VOK> ^ <CS=130,FS=130> / L &
41.09: <VOK> ^ <CS=130,FS=130> / L1 &
41.10: <VOK> ^ <CS=130,FS=130> / <VOK> &
42.00: L1 ^ <CS=130,FS=130> / & <VOK>
42.01: L1 ^ <CS=200,FS=200> / <OCLUDEF> &
las transiciones entre sordas continuas comienzan en la primera sorda
42.02: <SEG,CONT> ^ <TK1=-X/3,TK2=-X/3> / <SEG,CONT,X:=DR> &
      parallel transitions between unvoiced non-fricative and unvoiced begin
at the non-fricative (points are inserted, not modified)
 42.03: <SEG,VOICE,-FRIC> ^ <K1:=X,K2:=Y,TK1=DR/3,TK2=DR/3> / & <SEG,-VOICE,X:
=K1,Y:=K2>.
las fricativas sordas continuas provocan fricación en el comienzo de la
siguiente o la previa continua sonora
42.04: <SEG,VOICE,-FRIC> ^ <TAC=1> / <SEG,-VOICE,+FRIC,+CONT> &
42.05: <SEG,-VOICE,+CONT,+FRIC> ^ <TAC=-1> / <SEG,VOICE,-FRIC> &
      las transiciones entre vocal y nasal son lentas y tempranas
42.06: NY ^ <TF2=-X/3,TF3=-X/3,TFS=-X/3,FS=80> / <SEG,VOC,X:=DR> &
      para no reducir la duración percibida de la vocal las transiciones entre
vocal y nasal se adelantan
42.07: <SEG,VOC> ^ <TF2=0,TF3=-X/3,TFS=0,FS=80> / NY &
      contextual modification due to preceding
//  or following back vowel
42.08: NY ^ <F2=192,F3=136> / & <-BACK>
42.09: NY ^ <F2=192,F3=136> / <-BACK> &
      tras pausa se comienza en el objetivo
42.10: <-SEG> ^ <K1=W,K2=X,AK=Y> / & <SEG,W:=K1,X:=K2,Y:=AK>
43.00: <-SEG> ^ <FN=W> / & <SEG,W:=FN>
      copiamos los formantes del fonema a la pausa posterior
43.01: <-SEG> ^ <F1=W,F2=X,F3=Y> / & <SEG,W:=F1,X:=F2,Y:=F3>
43.02: <-SEG> ^ <TF1=DR/4, TF2=DR/4, TF3=DR/4, TF4=DR/4, TK1=DR/4, TK2=DR/4,
TFN=DR/4, TAK=DR/4>
      las transiciones abruptas en el silencio (para no oírlas)
 43.03: <-SEG> ^ <CS=200,TCS=DR/4,FS=200,TFS=DR/4>
43.04: <-SEG> ^ <A0=0> / <-SEG> &
      copiamos los 2 primeros formantes del contexto posterior
F1=348,F2=1745
43.05: R1 ^ <F1=76,F2=172> / & I
      F1=388,f2=1645
43.06: R1 ^ <F1=92,F2=164> / & E
      F1=469,f2=1482
43.07: R1 ^ <F1=116,F2=148> / & A
      F1=396,f2=1213
43.08: R1 ^ <F1=96,F2=120> / & O
      F1=340,F2=1111
43.09: R1 ^ <F1=72,F2=108> / & U
      F1=348,F2=1745
43.10: R ^ <F1=76,F2=172> / & I
      F1=388,F2=1645
```

44.00: R ^ <F1=92,F2=164> **/ & E**

precedida de í', é'.

44.01: R ^ <F1=76,F2=172> **/ I &**

44.02: R ^ <F1=92,F2=164> **/ E &**

modificamos el punto de F2 para insertar explosión

44.03: <D1,X:=F2> ^ <F1=44,F2:=X,TF2=DR+2> **/ & <+ROUND>**

colocamos la explosión al final de la oclusiva (la duración pudo ser modificada por reglas contextuales)

44.04: P ^ <TLAC=DR,2TLAC=DR-1,TLAN=DR-2>

44.05: T ^ <TLAC=DR,2TLAC=DR-1,TLAN=DR-2>

44.06: K ^ <TLAC=DR,2TLAC=DR-1,TLAN=DR-2>

44.07: K ^ <2LAC=15,TLAC=DR,2TLAC=DR-2,TLAN=DR-2> **/ & T**

la explosión de las oclusivas: transiciones bruscas de los formantes

44.08: <+VOK> ^ <TA0=-1,TST=-1,ST=28,CS=160,FS=160,TCS=-1,TFS=-1,TF1=-1,TF2=-1,TF3=-1,TF4=-1> **/ P &**

44.09: <+VOK> ^ <TA0=-1,TST=-1,ST=28,CS=160,FS=160,TCS=-1,TFS=-1,TF1=-1,TF2=-1,TF3=-1,TF4=-1> **/ T &**

44.10: <+VOK> ^ <TA0=0,TST=-1,ST=48,CS=130,FS=80,TCS=-1,TFS=-1,TF1=-1,TF2=-1,TF3=-1,TF4=-1> **/ K &**

45.00: R1 ^ <TA0=-1,TST=-1,ST=28,CS=160,FS=160,TCS=-1,TFS=-1,TF1=-1,TF2=-1,TF3=-1,TF4=-1> **/ P &**

45.01: R1 ^ <TA0=-1,TST=-1,ST=28,CS=160,FS=160,TCS=-1,TFS=-1,TF1=-1,TF2=-1,TF3=-1,TF4=-1> **/ B1 &**

45.02: R1 ^ <TA0=-1,TST=-1,ST=28,CS=160,FS=160,TCS=-1,TFS=-1,TF1=-1,TF2=-1,TF3=-1,TF4=-1> **/ D1 &**

adelantamos el arranque de A0 como en oclusivas. incrementamos la velocidad de transición

45.03: R1 ^ <TA0=0,TST=-1,ST=28,CS=160,FS=160,TCS=-1,TFS=-1,TF1=0,TF2=0,TF3=0,TF4=0> **/ K &**

45.04: L1 ^ <TA0=-1,TST=-1,ST=28,CS=160,FS=160,TCS=-1,TFS=-1,TF1=-1,TF2=-1,TF3=-1,TF4=-1> **/ P &**

45.05: L1 ^ <TA0=-1,TST=-1,ST=28,CS=160,FS=160,TCS=-1,TFS=-1,TF1=-1,TF2=-1,TF3=-1,TF4=-1> **/ B1 &**

45.06: L1 ^ <TA0=-1,TST=-1,ST=28,CS=160,FS=160,TCS=-1,TFS=-1,TF1=-1,TF2=-1,TF3=-1,TF4=-1> **/ G1 &**

Coarticulación de las oclusivas labiales delante de 'O' y 'U'

45.07: <+OCLUDEF,+ANT,-COR> ^ <F2=0,F3=Y-30,FS=130> **/ &** *<+VOK,+ROUND,X:=F2,Y:=F3>*

antes de 'I'

45.08: <+OCLUDEF,+ANT,-COR> ^ <F1=0,CS=130> **/ & <+VOK,+HIGH,X:=F1>**

45.09: I ^ <1F1=112,1TF1=-1> **/ <+OCLUDEF,+ANT,-COR> &**

45.10: I ^ <2F1=48,2TF1=1> **/ <+OCLUDEF,+ANT,-COR> &**

46.00: U ^ <1F1=112,1TF1=-1> **/ <+OCLUDEF,+ANT,-COR> &**

46.01: U ^ <2F1=48,2TF1=1> **/ <+OCLUDEF,+ANT,-COR> &**

la sonoridad en las oclusivas se consigue nasalmente

46.02: B1 ^ <A0=0,1AN=10,1TAN=0,LAN=0,LTAN=DR>

46.03: G1 ^ <A0=0,1AN=15,1TAN=0,LAN=0,LTAN=DR>

46.04: D1 ^ <A0=0,1AN=15,1TAN=0,LAN=0,LTAN=DR>

'B' 'D' F1=300. para 'E', 'O', 'A'

46.05: <+APRODEF,+ANT> ^ <F1=56> **/ <+VOK,-HIGH> &**

F1=300. para 'E', 'O', 'A'

46.06: J ^ <F1=56> **/ <+VOK,-HIGH> &**

'E', 'O' 'A', F1=300

46.07: <+OCLUDEF,+ANT,+VOICE> ^ <F1=56> / **<+VOK,-HIGH> &**
para 'B','D' SEGUIDOS DE á', é', ó'
46.08: <+OCLUDEF,+ANT,+VOICE> ^ <F1=56> / **&** **<+VOK,-HIGH>**
'P' y 'T' seguidas por 'E', 'O'
46.09: <+OCLUDEF,+ANT,-VOICE> ^ <F1=56> / **&** **<+VOK,-HIGH>**
'P' y 'T' tras 'E', 'O'
46.10: <+OCLUDEF,+ANT,-VOICE> ^ <F1=56> / **<+VOK,-HIGH> &**
'T' seguida por A F1=500
47.00: T ^ <F1=128> / **&** **A**
'T' tras 'a'
47.01: T ^ <F1=128> / **A &**
F2=1500 para 'T' seguida o precedida por 'A'
47.02: T ^ <F2=152> / **A &**
47.03: T ^ <F2=152> / **&** **A**
T detrás o delante de 'o', 'u'.F2=898
47.04: T ^ <F2=80> / **&** **<+VOK,+ROUND>**
47.05: T ^ <F2=80> / **<+VOK,+ROUND> &**
copiamos F2 de la siguiente vocal
47.06: G1 ^ <F2=X> / **&** **<+VOK,X:=F2>**
47.07: G1 ^ <F3=X> / **&** **<I,X:=F3>**
Coarticulación de 'K' y 'G1' delante o detrás de 'E', 'O', 'A'
47.08: <+OCLUDEF,-ANT,-COR> ^ <F1=56> / **<+VOK,-HIGH> &**
47.09: <+OCLUDEF,-ANT,-COR> ^ <F1=56> / **&** **<+VOK,-HIGH>**
Coarticulación de la velar 'K'
47.10: K ^ <F2=X+48,F3=X-30> / **<SEG,X:=F2> &**
48.00: K ^ <F2=X+48,F3=X-30> / **&** **<SEG,VOICE,X:=F2>**
estas reglas afectan a 'K' y 'G1'
48.01: <OCLUDEF,-ANT,F2>220> ^ <F2=220,F3=F2-78>
48.02: <OCLUDEF,-ANT,F3>160> ^ <F3=160>
48.03: <OCLUDEF,-ANT> ^ <K1=F2-80,K2=F3-48>
48.04: <OCLUDEF,+ANT,-VOICE> ^ <TK1=-3,TK2=-3> / **<+SEG,-FRIC> &**
aceleramos CS e incrementamos F1 para percibir G1 y L148.05: G1 ^ <F1=132> / **&** **L1**
48.06: G1 ^ <CS=200> / **&** **L1**
sonidos africados: copiamos los formantes
48.07: CH ^ <F1:=W,F2:=X,F3:=Y> / **&** **<+VOK,W:=F1,X:=F2,Y:=F3>**
48.08: J1 ^ <F1:=W,F2:=X,F3:=Y> / **&** **<+VOK,W:=F1,X:=F2,Y:=F3>**
48.09: <FRICDEF,+ANT> ^ <F1:=W,F2:=X,F3:=Y> / **&** **<W:=F1,X:=F2,Y:=F3>**
PONEMOS FRICACIÓN EN 'D' FINAL.
48.10: D ^ <TAC=0,AC=5> / **&** **<-SEG>**
HACEMOS REGLA GENERAL PARA FRICACIÓN DE FRICATIVAS
49.00: <FRICDEF> ^ <TAC=0,AC=10>
VAESS: Emociones
49.01: <X:=FD> ^ <FD=FD+X3-100>
*49.02: <X:=F0> ^ <F0=F0*X3/100+100-X3>*
Si incrementamos F0, debemos incrementar el ancho de banda
*49.03: <X:=B1> ^ <B1=B1+(X3-100)*40/256>*
49.04: <VOK,STRESS,X:=ST> ^ <ST:=X+(X1-100)/25,TST=-2>
49.05: <VOK,STRESS,ST<1> ^ <ST=1>
Todos los segmentos tendrán menor A0, incluso los acentuados si X1=0, mantiene el valor anterior de A0, si X1=100, lo decrementa.
*49.06: <SEG,X:=A0> ^ <A0=X*5/6>*
49.07: <VOK,STRESS,X:=A0> ^ <A0=X(600-X1)/500>*
valor por defecto FA=1600

```
49.08: <> ^ <FA:=80>
49.09: <X:=FA> ^ <FA=X*X5/100>
49.10:   ^ <NA:=X6*100/256>
50.00: <VOK,STRESS> ^ <DR=DR*X7/100>
50.01: <VOK,-STRESS> ^ <DR=DR*100/X7>
      un valor alto de X8 incrementará OQ porque se incrementan RK y RG
50.02: <X:=RG> ^ <RG=X*(143-43*X8/100)/100>
50.03: <RG>165> ^ <RG=165>
50.04: <RG<80> ^ <RG=80>
50.05: <X:=RK> ^ <RK=X*(24+98*X8/100-(23*X8*X8/100)/100)/100>
50.06: <RK>60> ^ <RK=60>
50.07: <RK<20> ^ <RK=20>
50.09: <VOK,+STRESS> ^ <D:=X9>
50.10: <VOK,+STRESS,D>50> ^ <F0=250> / <-SEG> <SEG,-STRESS>(,) &
50.10: <VOK,+STRESS,D<15> ^ <DR=DR*(X7+40)/100> / & <SEG,-STRESS>(,) <-SEG>
51.00 SPEAK
51.01 FIN
```

---

[1] Chomsky está muy interesado en las adecuaciones descriptiva y sobre todo explicativa, el realismo de la computación mental, el innatismo del lenguaje humano y su aprendizaje, y la independencia entre el sistema lingüístico-cognitivo y el sistema racional

[2] Evaluación realizada sobre más de 700.000 palabras procedentes del suplemento cultural del diario ABC

[3] Sin incluir signos de puntuación

[4] Piénsese que si a cada palabra del lexicón se le asigna una etiqueta unívoca, la ambigüedad es 0.

[5] Para las palabras fuera de vocabulario se emplearon los rasgos: comienza por mayúscula, contiene guión, contiene números, sufijo de 3 letras (lo cual parece poco en castellano), etiqueta ambigua de la palabra que desambiguar y de la palabra a su derecha y etiqueta desambiguada a la izquierda

[6] Obsérvese que los sistemas *TnT* y *JMX* no permiten extensión de diccionarios, por eso no se dan datos sobre sus experimentos

[7] Entroncando con la sintaxis X-barra y el programa minimalista podríamos hablar de especificadores, núcleo y complementos (*N. Chomski* 1995 "The minimalist program" *MIT Press*)

[8] Las gramáticas de unificación con rasgos complejos recursivos posee potencia equivalente al de una máquina de Turing, resultando un modelo posiblemente excesivo para el procesamiento de lenguaje natural

[9] http://www.ling.mq.edu.au/rmannell/sph307/ lecture8_psychacoustics/chapter2.html

[10] **Estilización**: extraer la secuencia mínima de tramos rectos de F0 que son percibidos de la misma manera que la secuencia o curva real (*P. Mertens & C. Allessandro* 1995). Ejemplos en castellano se presentan en (*E. López Gonzalo* 1993) y (*J. M. Garrido* 1991). Algunos métodos (por su intrínseco suavizado) no necesitan estilización.

[11] Obsérvese que aunque se use el error cuadrático medio mínimo se trabaja en log F0.

[12] Entre estas alteraciones podemos destacar: conductividad de la piel, la expresión facial o vocal, la dilatación de las pupilas, la frecuencia de respiración, la tensión sanguínea o el pulso (*J. Healey & R. Picard* 1998).

[13] Son las llamadas Big Six: alegría, tristeza, enfado, sorpresa, miedo y repugnancia (*R. Cornelius* 2000)

[14] Los modelos prosódicos no se encontraban disponibles por la privacidad de un producto comercial como DECTALK.

[15] Es importante el orden en la evaluación: primero deben aparecer las respuestas libres y luego las forzadas, y primero deben aparecer los textos neutros y luego los cargados con contenido emotivo.

[16] los parámetros sílaba inicial o final han sido definidos como se describió antes

[17] Preguntados por las razones de la calificación de "pobre calidad de voz", los oyentes que la asignaron comentaron que habían experimentado dificultades para identificar la emoción expresada. Sin embargo, al comparar sus resultados posteriores en identificación de emociones con los de otros sujetos, su nivel de acierto era incluso superior, aunque sin significancia estadística.

[18] Sólo se dispone de 2 grabaciones de voz neutra en vez de 3.

[19] Obsérvese que no es el texto el que confiere carácter alegre a la frase, dado que cualquiera de nuestras emociones es compatible con el texto.